NOVEL METHODS FOR HISTORICAL INDUS DOCUMENT UNDERSTANDING

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY (Ph.D)

In

COMPUTER SCIENCE

(UNDER THE FACULTY OF COMPUTER SCIENCE AND TECHNOLOGY)

By

KAVITHA A.S

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE
UNIVERSITY OF MYSORE, MANASAGANGOTRI
MYSURU-570 006, INDIA

APRIL 2017

UNIVERSITY OF MYSORE

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE

MANASAGANGOTHRI

MYSURU – 570 006, INDIA

DECLARATION

I hereby declare that the entire work embodied in this Doctoral thesis has been carried out by me at the Department of Studies in Computer Science, University of Mysore, Mysore, under the supervision of **Prof.G Hemantha Kumar**. This thesis has not been submitted in part or full for the award of any diploma or degree of this or any other University.

Kavitha A.S
Research Scholar
Dept. of Studies in Computer Science
University of Mysore
Manasagangothri
Mysuru – 570 006, INDIA.

UNIVERSITY OF MYSORE

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE

MANASAGANGOTHRI

MYSURU – 570 006, INDIA

CERTIFICATE

This is to certify that **Mrs. Kavitha A.S** has worked under my supervision for her Ph.D. thesis entitled "**Novel Methods for Historical Indus Document Understanding**". I also certify that the work is original and has not been submitted to any other university wholly or in part for any other degree.

Dr. G Hemantha Kumar
Professor and Research
Supervisor
Department of Studies in
Computer Science
University of Mysore
Manasagangothri
Mysuru – 570 006
INDIA.

Prof. D.S.Guru
Chairman
Department of Studies in
Computer Science
University of Mysore
Manasagangothri
Mysuru – 570 006
INDIA.

DEDICATED

TO

MY FAMILY AND

TEACHERS

Acknowledgements

First and foremost praises and thanks to God, Lord Ganapathi for his showers of blessings throughout my research work.

I wish to thank the following people for ease of my work during research.

It is my immense pleasure to express sincerely, thanks and gratitude to my research guide, Prof. Dr. G. Hemantha Kumar for his invaluable guidance constantly throughout entire research work. I would also thank him a lot for the accommodation during my stay at Mysore. I would also thank him without his moral help I would not managed studies, family and profession. I am very fortunate to do my research with such a good person. His patience, knowledge helped me a lot to reach the goal of research.

I wish to express my gratitude to my teachers Dr. P. Nagabhushan, Dr. D.S.Guru, Dr. Lalitha Rangarajan of Department of Studies in Computer Science who laid the foundation for taking up the research work. I would like to thank them for laying strong foundation to take up research work during the course work.

I thank my teacher Dr. Suresha, DOS in Computer Science who put lot of effort in getting my course work certificate from university. I thank him very much considering my request. I thank all my teachers Dr.H.S.Nagendraswamy, Smt. L. Hamsaveni and Dr.Sharada and other faculty members of Department of Studies in Computer Science for their suggestions, wishes and moral support during my research endeavour.

I would like to take this opportunity to express my sincere special thanks to my teacher, colleague and friend Dr. P. Shivakumara, University of Malaya, Malaysia. It would not be mistake if I tell he is the person who taught me from the word image. He had supported me in all aspects without looking time. I would like to say my endless thanks to him. I thank him for his invaluable technical support given to me during the entire research though he is far.

I would like to thank Vice Chancellor, Registrar and all the administrative staff of University of Mysore for their help and support for permitting me to take up research work in the Department of Studies in Computer Science, Manasagangothri, and Mysore.

I also express my thanks to all the teaching and non teaching staff of Department of Studies in Computer Science, Manasagangotri for their help and support.

I will be failing if I don’t acknowledge Dr. Basavanna, Government College, and Mandya who taught me Latex. I would like to convey my thanks to him. I would also thank the staff of Archaeology department Mysore for their help in data collection. I also wish to thank Dr. Chetan H.K., Maharaja Institute of Technology and Dr. Vinay for their support in early stage of research.

I place a record of thanks to my research friends Dr. K.S. Manjunath, Kouser, Sudha, Chethan, Sheela, Rashmi , Naveen,Dr. Mohammed Javed, Preethi, Manjula and all other friends for their constructive support during the Ph.D work.

This thesis would not been possible without the support of my family members. I thank my father Sri. A.S.Shivegowda and mother Smt. Rathna who has supported by taking care of the family. Work has not been initiated without the support of my husband Dr. M.N.Sanjaya., who always stood behind my success. Last but never the least, I extend my heartfelt thanks to him. I also thank my little son, Supreeth S for his smile, love and affection in my life who is playing role of relief and inspired me to face challenges. I also wish to thank my in-laws for their moral support and my co-brothers Ravi and Anand for their timely support.

Finally, I thank all my friends and relatives who directly or indirectly helped me during the research work.

Kavitha A.S

Abstract

Document image understanding is a successful application in the field of document image processing and pattern recognition as we can see Optical Character Recognizer (OCR) for several scripts in the market which gives more than 90% recognition rate for plain and homogeneous background document images. Due to rapid change in technology, instead of capturing images through conventional scanner which is very slow and not user friendly, images are captured through camera with high resolution. As a result, images captured by camera get affected by several causes, such as blur due to camera moments, noise due to open environment, perspective distortion due to non-orthogonal direction of the camera etc. Therefore, the methods developed for scanned images may not work for the images captured by camera without modifications.

To overcome the problems of camera based images, there are the methods developed for correcting alignment of images, increasing contrast of text information from the background etc. However, the developed methods still work based on the fact that images have plain background with high contrast and printed text. These methods may not work well for historical document images where we can see severe degradation of the printed text as well handwritten texts due to ageing, folding, material that used for printing/writing, variations in handwriting style and the way they preserve the documents. Therefore, in this way, historical document image processing has become the hottest research topic for the document analysis community.

There are methods proposed for finding solution to degradation of the handwritten text, printed text, low contrast and distorted text etc. But the methods still require homogeneous background images for achieving better results. As a result, these methods may not be suitable for the historical documents such as Indus document where it can be seen complex background with lots of variation in writing style, variation in resolution, shadow effect, touching characters to characters, touching characters to non-text, variety of shapes and broken characters etc. This is because during Indus document period, people used rocks, hard material, walls, palm leaves, cloths etc with sticks for writing but not like modern pen with ink and plain paper. Therefore, understanding Indus document requires special attention and new methods. Furthermore, it is essential to preserve the history for future and non-

availability of experts to translate the old script into known script. Thus we consider Indus document understanding for our research as it is challenging and useful to the society.

In this thesis, to find solution to Indus document images, we divide the whole problem into four sub-issues that are Indus script identification from other scripts to choose Indus document for processing further, text lines segmentation from non-text in the Indus document, character segmentation from text lines and character recognition to understand the Indus document.

It is true that when we see the nature of Indus document, one can expect components with more cursive and less straightness compared to normal documents such as English, Kannada, Tamil, Telugu and Malayalam etc where we can see clear shape of the components. This is due to complexity of the background and the presence of animal picture which connect to text components in case of Indus document. This observation motivates us to propose a method for studying cursiveness and straightness of the components based on morphological operation to group the disconnected components as one and region growing for merging neighbour components. However, this method gives more false positives due to region growing and hence poor classification rate. Therefore, to improve the classification rate, we introduce skewness property for extracting the same straight and cursiveness of the components. However, this method is limited to few numbers of classes. Therefore, we propose a new method which extracts spatial relationship between corners, end, junction, intersection points to study structure of the components. Variances of distance between dominant points are considered for script identification.

To segment text line from the Indus script document image, we propose a new method based on nearest neighbour criterion. It is known, a text component exhibit high contrast compared to non-text in Indus document. With this clue, we propose a new method which is combination of Sobel and Laplacian for separating text and non-text components. Then the proposed method use boundary growing based nearest neighbour criterion to segment text lines. This idea works well because the space between text components is lesser than the space between text lines.

To segment character from text lines, we propose a method based on watershed model because watershed model give direction according to water flow. It works well for touching characters as it provides information about water storage called catchment basis area if character is touching else clear path between the characters. The proposed method studies ridges given by watershed model and catchment basin for final character segmentation.

To recognize the characters, to understand the Indus document image, we propose a new descriptor called Histogram Oriented Tangent Descriptor. The proposed method divides the whole character image into equal number of blocks of variable size to convert to standard size. For each pixel in the block, we estimate tangent angle, slope. Further, the angle information in encoded into number of bins according to angle information as in Histogram Oriented Gradients descriptor (HOG) in new way for extracting shape of the character components.

Finally, the proposed algorithms are evaluated using standard measures such as classification rate for script identification, recall, precision, F-measure for text line and character segmentation and recognition rate for character recognition and standard databases. To show the effectiveness of proposed method, we implement well-known existing methods for comparative study.

Contents

Chapter 1: Introduction

Chapter 2: Literature review

Chapter 3: Indus script identification

Chapter 4: Text line segmentation

Chapter 5: Character segmentation

Chapter 6: Character recognition

Chapter 7: Conclusion and future work

List of Tables

List of Figures

Chapter 1

Introduction

It is true that one of the successful system in the field of document image analysis is Optical Character Recognizer (OCR) engine which is used for converting documents to digital form which in turn used for several applications, such as understanding documents, automaton of office tasks with the aim of paperless office, converting one script to another script as the people can understand documents by reading their own languages etc. In addition, the OCR is useful in preserving the document information for future study without losing information. In this way, developing OCR gained lots of interest of the researchers in the field of document image analysis. The general architecture of OCR can be seen in Fig. 1.1 where it acquires image through devices, such as scanner, camera etc and it extracts features to represent the image, which can be used further for text line segmentation, character segmentation from text lines and character recognition. However, it is noted that the successful OCR engine works well for the document scanned by high resolution with plain background. There are other documents for which the conventional OCR does perform well, such as historical documents where one can expect low resolution due to aging, distortion due to fold etc, degraded documents which contains low quality, camera based documents which contains perspective distortion due to angle variation during capturing an image etc. Therefore, there is an immense demand for developing robust OCR for achieving recognition results for the above mentioned documents.

There are methods developed to overcome the above causes of such documents in literature. However, most of the methods focus on the documents affected particular adverse factor such as degradation, distortion, low quality, low resolution etc. There are documents which affected by multiple adverse factors, such as low quality, low resolution, complex foreground due to writing style with the wooden sticks, complex background due to use of palm leaves, rock for writing etc, namely, Indus documents. In this work, we focus on developing OCR for understanding Indus document which affected by multiple diverse factors. Besides, to the best our knowledge, Indus document recognition and understanding is considered as open issue in the field of document image analysis and pattern recognition.

The chapter is organized as follows. Section 1.1 presents history of document imaging which focus on evolution of document image processing, Section 1.2 describes history of Indus document which focus on degraded historical documents recognition, Section 1.3 identifies the challenges of Indus document recognition, Section 1.4 proposes a plan for tackling challenges of Indus document image recognition, the contributions are listed in Section 1.5 and lastly, organization of the thesis is presented in Section 1.6.

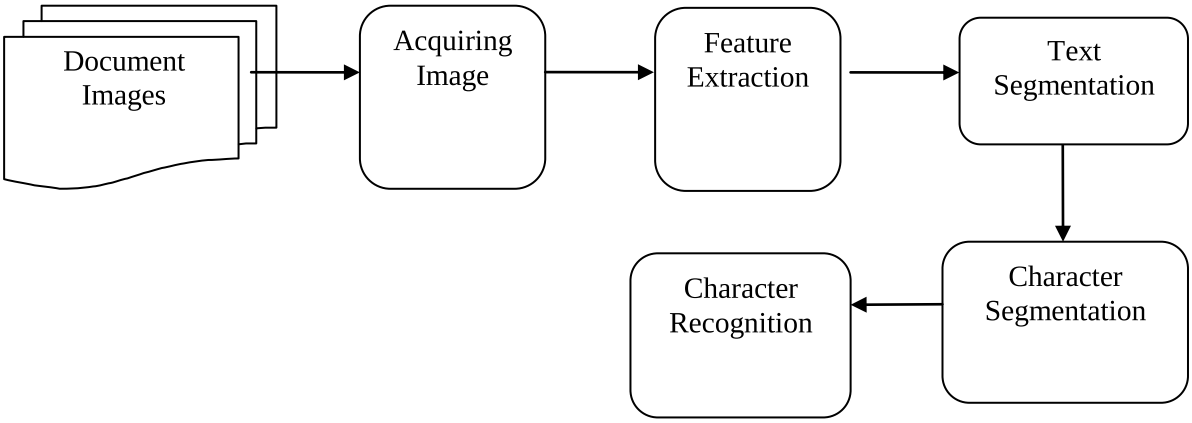


Fig. 1.1. General OCR Architecture

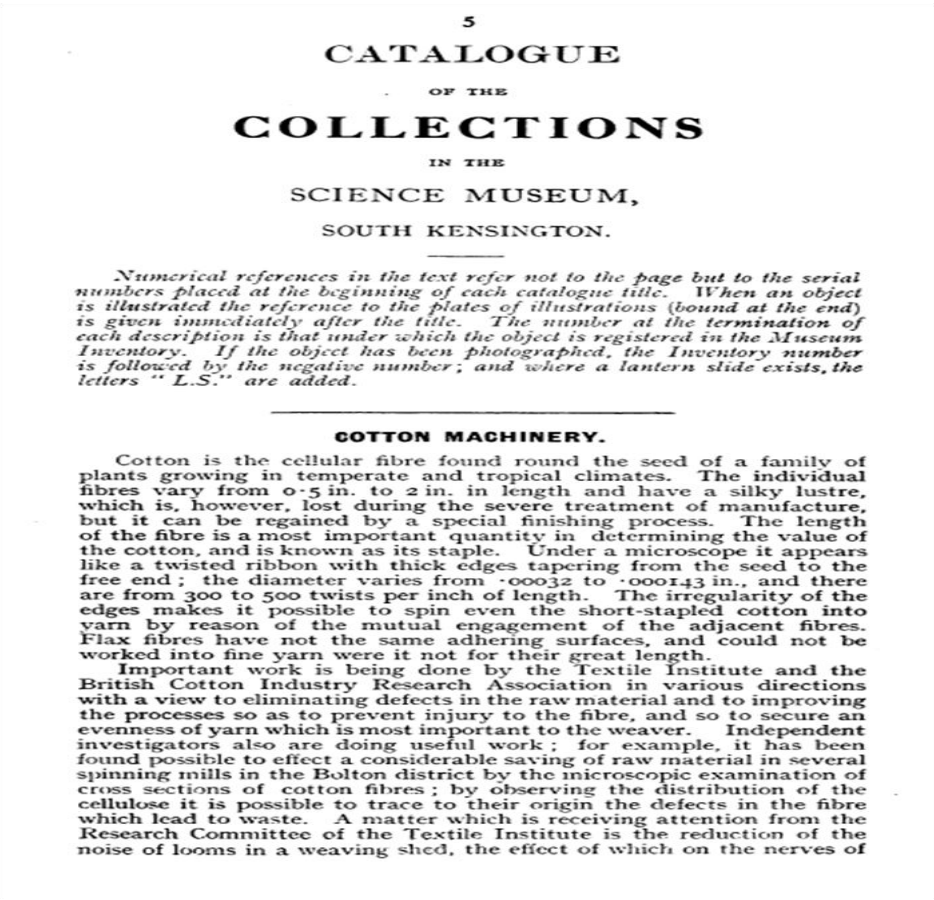
1.1 History of Document Imaging

As discussed in previous section, it is noted that document images can be classified broadly as scanned document images, camera based document images, degraded images and Historical-degraded images as shown in Fig. 1.2 where we can see the difference between the documents in terms of causes and difficulties in understanding the documents. It is observed from Fig. 1.2(a) that the documents appear simple because it has plain background and binary image. For these images, the conventional OCR works well. The document images shown in Fig. 1.2(b) consist of different colored background and foreground compared the documents in Fig. 1.2(a). This makes challenging to recognize the documents for the conventional OCR. The documents shown in Fig. 1.2(c) are degraded documents which suffer from low quality, low resolution and distortion due to problem in capturing devices. It is noted that despite degradations, the documents contain plain background. Furthermore, it is found from Fig. 1.2(d) that historical documents have complex background and the documents suffer from low quality due to aging and lack preservation.

There are approaches proposed in literature to solve the above mentioned problems. For example, detection and correction of skew in scanned document is proposed by Rajeev et al., 2015. This method explores Principal component analysis and Hough transform used for skew detection. This approach is useful when the scanned document is skewed due to fault at devices. If the document is skewed, this step affects subsequent steps, such as text line segmentation, page layout analysis and document recognition. Zhenwen et al., 2014 proposed a method for cleaning scanned documents because usually while scanning, devices introduce noise. The method recognizes the characters by taking document patches as input, which explores probabilistic model. From the above discussion, it can be concluded that there are methods which handles the issues of scanned document images. Since most of the images are scanned by high resolution scanner, the scanned documents suffer from noise and skew. These issues are addressed well in the literature to recognize the scanned documents. However, the same methods cannot be used directly to understand the document captured by camera devices due to distortions.

Similarly, for recognizing documents captured by camera devices which generally suffer from illumination effect, distortion, blur etc, document image retrieval system using local Features is proposed by Dang et al., 2015. Scale and Rotation Invariant Features (SRIF) are computed based on geometrical constraints between pairs of nearest points. SRIF extracts centroids of word connected components. However the method does not perform well for the images with complex background. Jagannathan et al., 2005, proposed a method for identifies corners within the bounding quadrilateral to identify the points. It identifies horizontal and vertical lines for perspective correction of text. This shows that the methods have been developed to overcome the problems of camera based images especially perspective distortion caused by different angle, rotations. However, these methods may not work well for degraded document images because degraded document images suffer low quality.

Restoration of degraded documents using image binarization technique is proposed by Kavya et al, 2015. The method produces the contrast map from the degraded document images and then the combination of local gradient and contrast for binarizing the degraded image. Karthika et al, 2014, proposed a method for binarization using bit map slicing to restore text pixels from degraded images. The method uses edge pixels detected by Laplacian which produces edge map. Kai et al, 2015, proposed a method for page segmentation based on unsupervised learning from degraded document images. The Method considers page segmentation as a pixel labeling problem, i.e., each pixel is classified as periphery, background, text block, or decoration. Similarly, segmentation-free pattern spotting in degraded document image is proposed by Sovann et al, 2015.The system includes a powerful patch-based framework, the bag of visual word model with an offline sliding window mechanism to avoid heavy computational burden during the retrieval process. This discussion shows that the developed methods work well for degraded document images to achieve better recognition. However, the methods are limited to plain background images. Therefore, the method may not work well for the degraded-historical document images, namely, Indus document images where one can see multiple adverse factors, such as low quality, low resolution due to aging, complex background due to rock palm leaves, wall etc, complex foreground due to stick and different writing style. Hence, Indus document recognition and understanding receives great attentions of the researchers.



5

CATALOGUE

OF THE

COLLECTIONS

IN THE

SCIENCE MUSEUM,

SOUTH KENSINGTON.

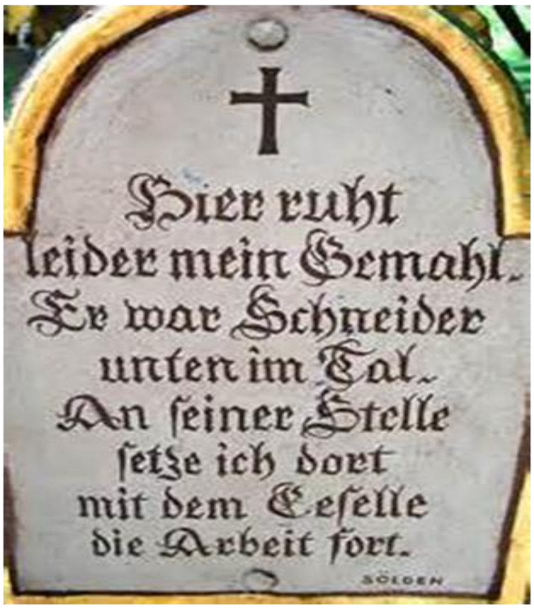
Numerical references in the text refer not to the page but to the serial numbers placed at the beginning of each catalogue title. When an object is illustrated the reference to the plates of illustrations (bound at the end) is given immediately after the title. The number at the termination of each description is that under which the object is registered in the Museum Inventory. If the object has been photographed, the Inventory number is followed by the negative number; and where a lantern slide exists, the letters "L.S." are added.

COTTON MACHINERY.

Cotton is the cellular fibre found round the seed of a family of plants growing in temperate and tropical climates. The individual fibres vary from 0·5 in. to 2 in. in length and have a silky lustre, which is, however, lost during the severe treatment of manufacture, but it can be regained by a special finishing process. The length of the fibre is a most important quantity in determining the value of the cotton, and is known as its staple. Under a microscope it appears like a twisted ribbon with thick edges tapering from the seed to the free end; the diameter varies from ·00032 to ·000143 in., and there are from 300 to 500 twists per inch of length. The irregularity of the edges makes it possible to spin even the short-stapled cotton into yarn by reason of the mutual engagement of the adjacent fibres. Flax fibres have not the same adhering surfaces, and could not be worked into fine yarn were it not for their great length.

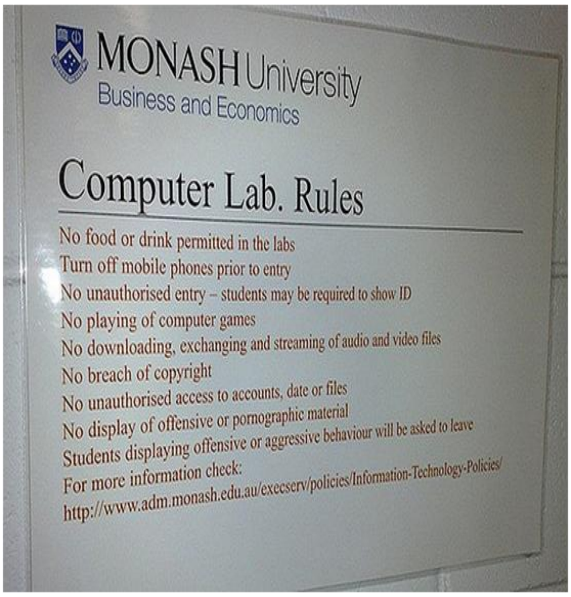
Important work is being done by the Textile Institute and the British Cotton Industry Research Association in various directions with a view to eliminating defects in the raw material and to improving the processes so as to prevent injury to the fibre, and so to secure an evenness of yarn which is most important to the weaver. Independent investigators also are doing useful work; for example, it has been found possible to effect a considerable saving of raw material in several spinning mills in the Bolton district by the microscopic examination of cross sections of cotton fibres; by observing the distribution of the cellulose it is possible to trace to their origin the defects in the fibre which lead to waste. A matter which is receiving attention from the Research Committee of the Textile Institute is the reduction of the noise of looms in a weaving shed, the effect of which on the nerves of

(b) Sample of Scanned document Images



Hier ruht
leider mein Gemahl.
Er war Schneider
unten im Tal.
An ſeiner Stelle
ſetze ich dort
mit dem Geſelle
die Arbeit fort.

SÖLDEN

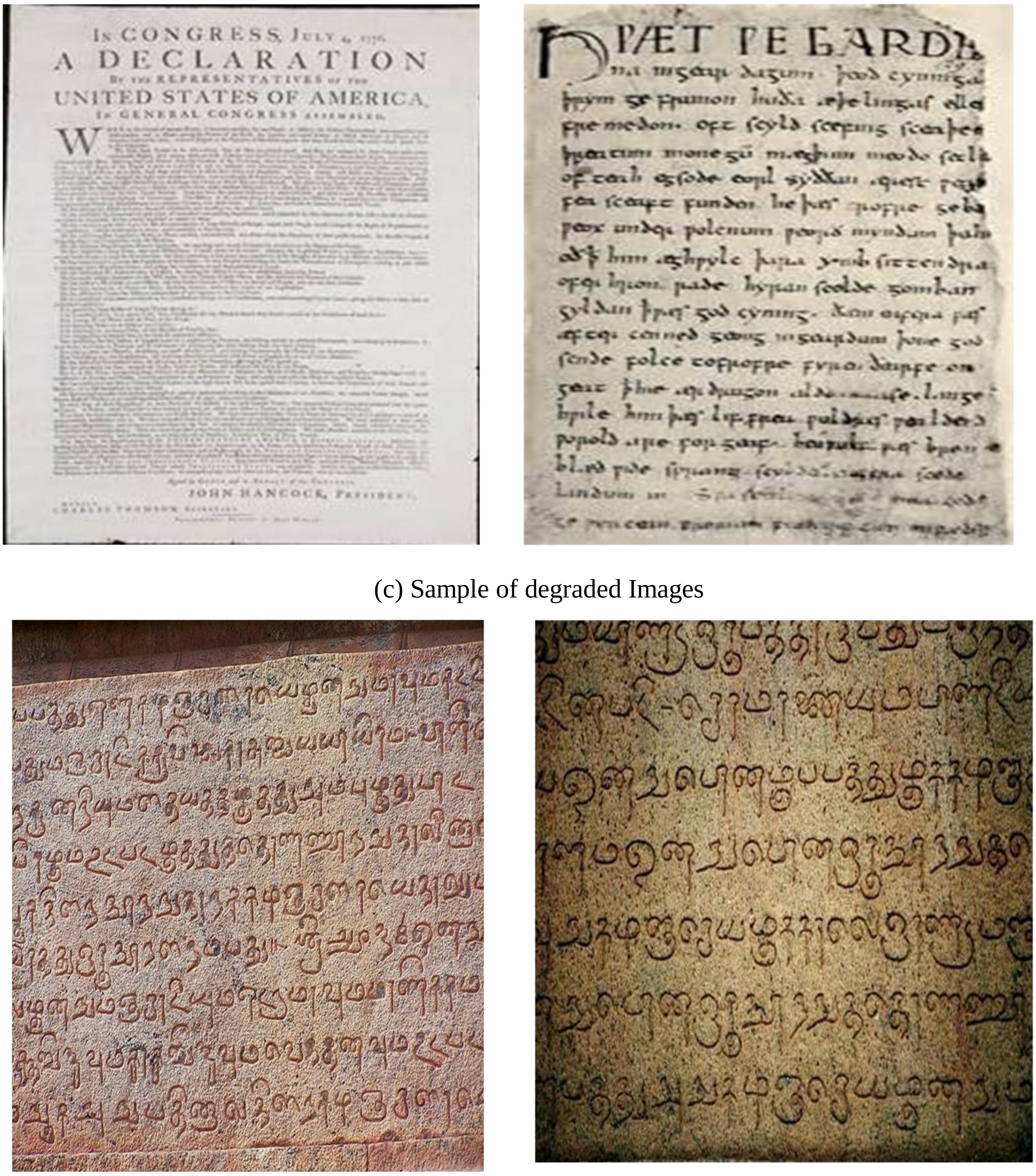


MONASH University
Business and Economics

Computer Lab. Rules

No food or drink permitted in the labs
Turn off mobile phones prior to entry
No unauthorised entry – students may be required to show ID
No playing of computer games
No downloading, exchanging and streaming of audio and video files
No breach of copyright
No unauthorised access to accounts, date or files
No display of offensive or pornographic material
Students displaying offensive or aggressive behaviour will be asked to leave
For more information check:
http://www.adm.monash.edu.au/execserv/policies/Information-Technology-Policies/

(a) Samples of Camera based Images


IN CONGRESS, JULY 4, 1776.
A DECLARATION
BY THE REPRESENTATIVES OF THE
UNITED STATES OF AMERICA,
IN GENERAL CONGRESS ASSEMBLED.
JOHN HANCOCK, PRESIDENT.
HWÆT WE GARDE

(c) Sample of degraded Images

(d) Samples of Historical scripts

Fig. 1.2. Evolution of document imaging.

1.2 History of Indus Document Images

Indus script is the photographic script of third millennium BC found in the Valley of Indus River. According to survey, the script was discovered in 1875. The Indus Script is the writing system developed by the Indus Valley Civilization and it is the earliest form of writing. Indus scripts consists set of symbols engraved during Indus valley civilization. Symbolic text like script engraved on some material like rock, clay etc found to be in the form of seals. These scripts engraved during period between 3500 and 1900 BC. Many cities existed during 3'rd BC. Remains of cities during that period are found during excavation by researchers and archeologists. Cities encompassed most of Baluchistan to some of Indian states like Haryana, Gujarat, Rajasthan, and Punjab.

Indus script also known as Harappa script found in the remains of such cities. Mohenjo-Daro and Harappa are the major cities during that period. Mohenjo-Daro is the city of the same period, located in Sindh province of Pakistan. These scripts show existence of civilization during that period so now called as, Indus civilization. During Indus civilization people unaware of tools for writing like pen, pencils. They used hard materials for carving. These carvings are carved on hard materials like stones, copper, clay etc. These carved symbol like text used for communication in business purpose. Most inscriptions are short and scripts written are unstructured. Characters are pictorial and include many signs. Approximately about 4000 such inscriptions found in various parts of the world. Texts found on the inscriptions are also pictorial in nature. These texts are also found on seals, miniatures and copper tablets. Indus scripts in seal form used for communication purpose since the people are unaware of languages during that period. It is clear that Indus scripts are earlier to Brahmi scripts. Fig 1.3 depicts the picture of different scripts evolved from Brahmi.

Sample images of evolution of historical documents can be seen in Fig. 1.4(a)-(f), respectively, Brahmi, Devanagari, Gupta Script, Gurumukhi Script, Script of Kalinga period, Kannada Script of Kadamba period. Fig. 1.4 shows that the document are complex to understand the content of the images compared to scanned, camera based and degraded document images because of complex background, low quality and different writing style. It is noted that Indus scripts was used about 1500 years earlier to Brahmi era. Hence, it is confirmed that Indus document are still complex than the above documents for recognition. Samples of some of Indus images shown in Fig 1.5 where we can notice the difference between Indus documents and the scanned, camera and degraded documents. Fig. 15 shows that the Indus document images generally have animal like picture, such as Unicorn, Elephant, Bull and Rhinoceros shown in Fig. 15(a)-(f), respectively. These inscriptions give hints about different people, traders, religious etc. As a result, the conventional methods which developed for scanned, camera based and degraded document images may not work well.

There are methods developed for recognizing the documents, such as epigraphically document like Indus in literature. For instance, Soumya et al., 2014, proposed a method for recognizing Kannada Epigraphically image. The method focus on Ancient Kannada scripts of Ashoka and Hoysala periods. The method basically extracts statistical features such as Mean, Variance, Standard Deviation, Kurtosis, Skewness, Homogeneity, Contrast, Correlation, Energy, and Coarseness for recognizing characters of documents. Similarly, to enhance the poor quality character images of epigraphically document, Rmanoorubini et al., 2014, proposed a method which uses morphological operation to remove degradations in the background of character images. Gangamma et al., 2012 proposed a method for prediction of era for epigraphically scripts. This method explores Curvelet Transform based approach for Classification of epigraphically scripts into various eras. This method involves Fast Discrete Curvelet Transform (FDCT) for prediction.

It is noted from the above discussion that there are methods which focus on epigraphically documents but not Indus documents which is older than epigraphically documents. Therefore, the methods developed for epigraphically documents may not work well for the Indus documents because Indus documents are more complex than epigraphically documents in terms of background, writing style and material which used for writing Indus text.

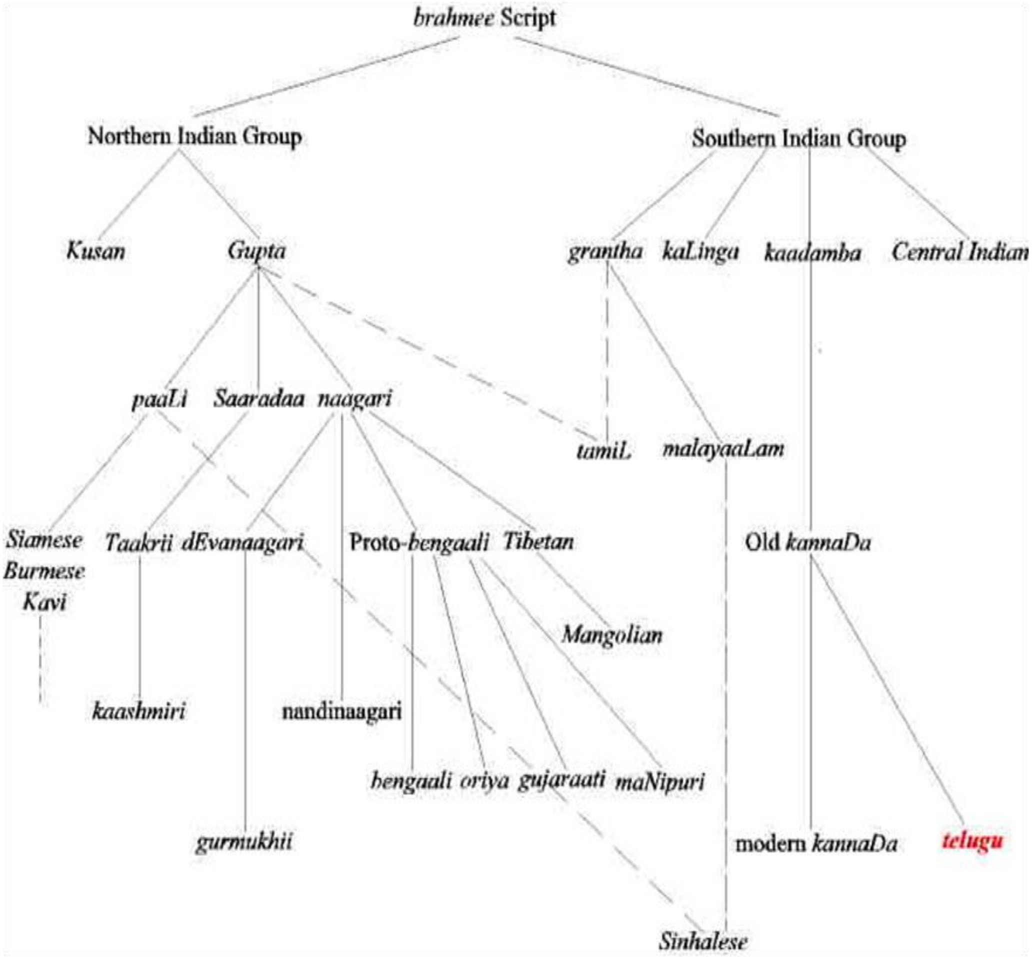
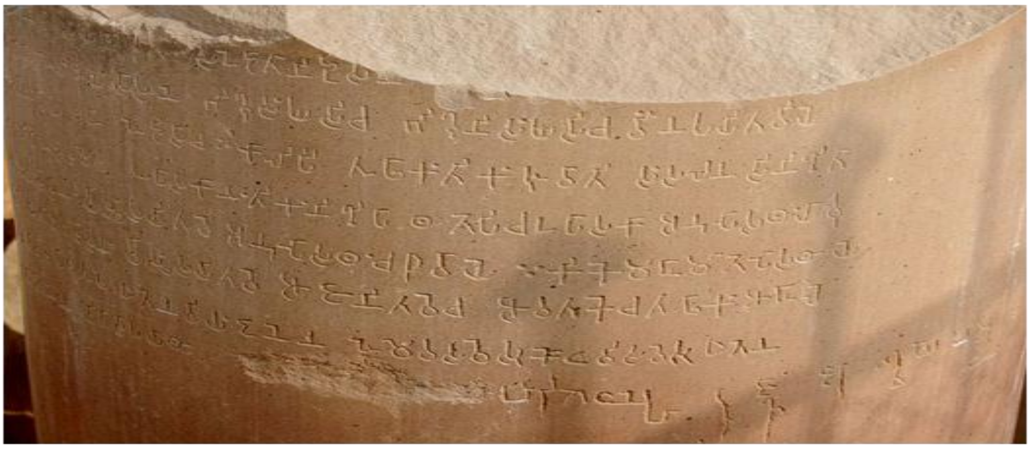
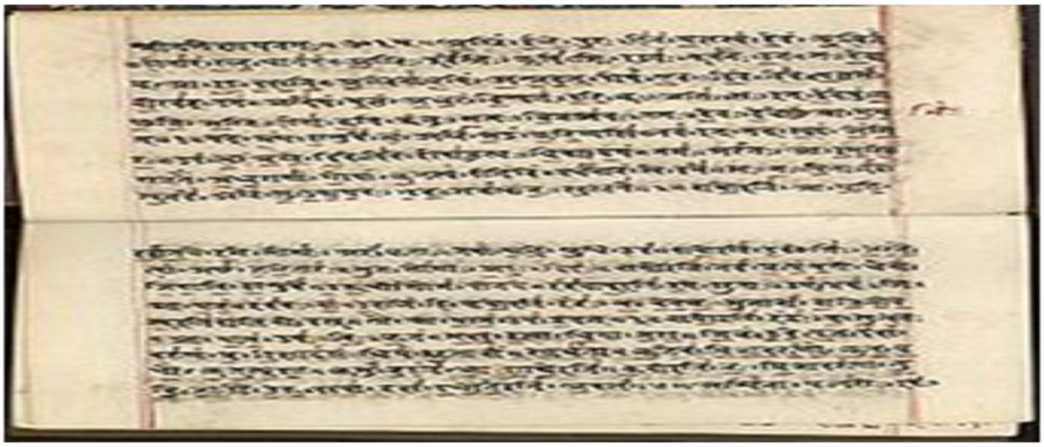


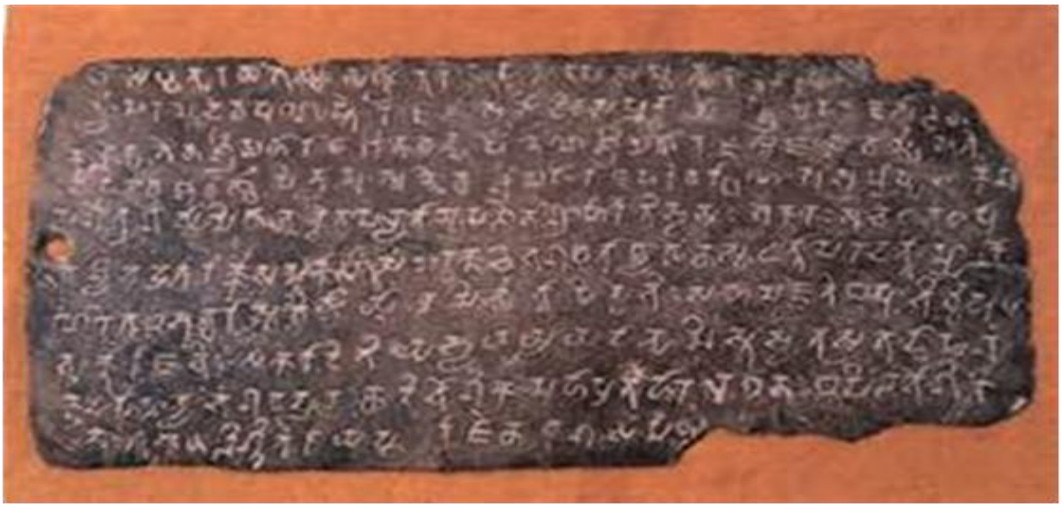
Fig. 1.3. Scripts descended from Brahmi



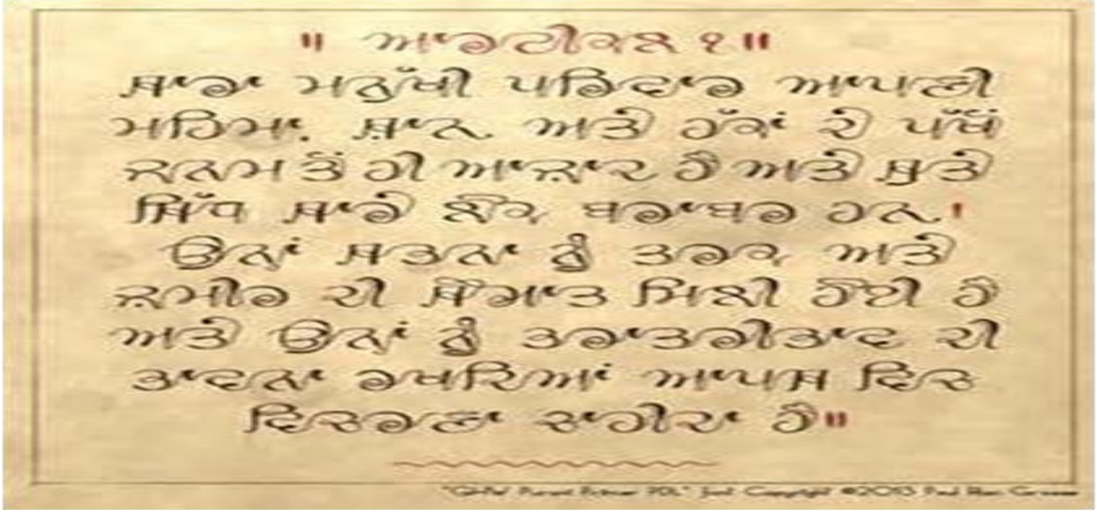
(a) Brahmi Script



(b) Devanagari Script



(c)Gupta Script



॥ ਆਰਟੀਕਲ ੧ ॥

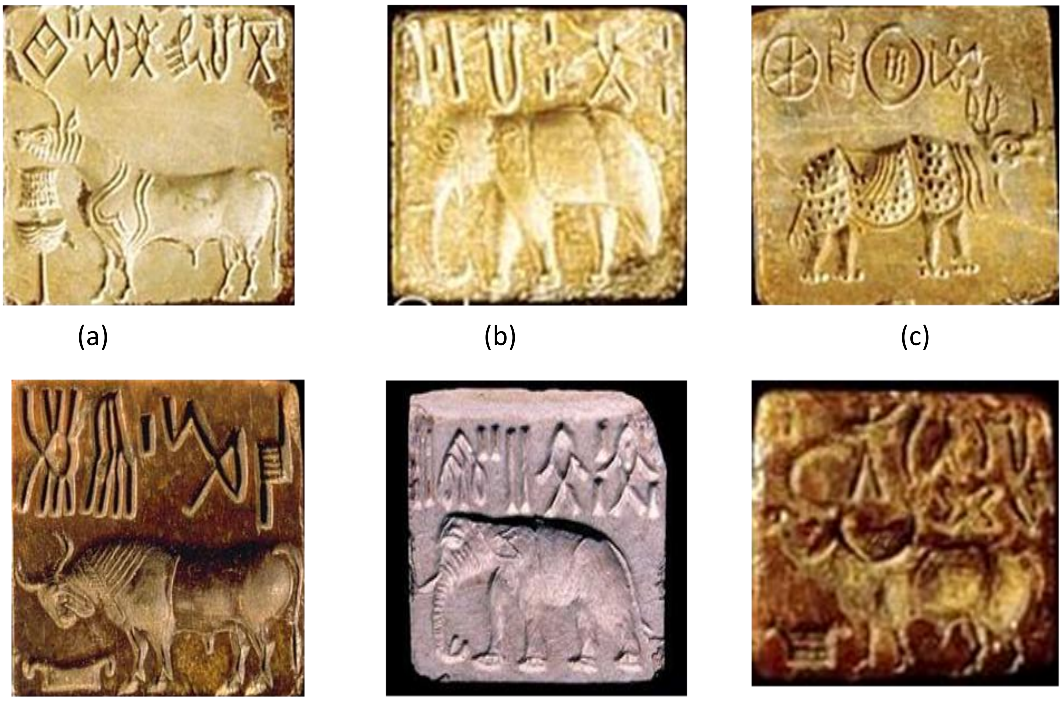
ਸਾਰੇ ਮਨੁੱਖੀ ਪਰਿਵਾਰ ਆਪਣੀ
ਮਹਿਮਾ, ਸ਼ਾਨ ਅਤੇ ਹੱਕਾਂ ਦੇ ਪੱਖੋਂ
ਜਨਮ ਤੋਂ ਹੀ ਆਜ਼ਾਦ ਹੈ ਅਤੇ ਸੁਤੇ
ਸਿੱਧ ਸਾਰੇ ਲੋਕ ਬਰਾਬਰ ਹਨ।
ਉਨ੍ਹਾਂ ਸਭਨਾਂ ਨੂੰ ਤਰਕ ਅਤੇ
ਜ਼ਮੀਰ ਦੀ ਸੌਗਾਤ ਮਿਲੀ ਹੋਈ ਹੈ
ਅਤੇ ਉਨ੍ਹਾਂ ਨੂੰ ਭਰਾਤਰੀਭਾਵ ਦੀ
ਭਾਵਨਾ ਰਖਦਿਆਂ ਆਪਸ ਵਿਚ
ਵਿਚਰਨਾ ਚਾਹੀਦਾ ਹੈ॥

(d) Gurumukhi Script

(e) Script of Kalinga period

(f) Kannada Script of Kadamba period

Fig. 1.4. Sample images of Scripts descended from Brahmi



(a) (b) (c)

(d) (e) (f)

Fig. 1.5. Sample images of Indus documents

1.3 Motivation

As noted from the discussion presented in previous sections, OCR for Indus document is still considered as open challenge for the researchers in the field of document image analysis. In addition, since the Indus documents refer history of the society and civilization and experts to interpret the documents are disappearing, it is necessary to preserve the documents content for our future study. Therefore, converting an Indus document to digital document is useful for the society. Besides, according to archeology department in Mysore, it is hard for them to annotate manually to interpret in modern language like English and Kannada because this requires huge manual effort and time. This factor justifies the importance of developing OCR for Indus document.

On top of usefulness, Indus document images are challenging to develop OCR for the researchers because of the following factors. The following are the additional challenges which are different from common challenges exists in case of Indus documents, such as font variations, distortion, noise etc.

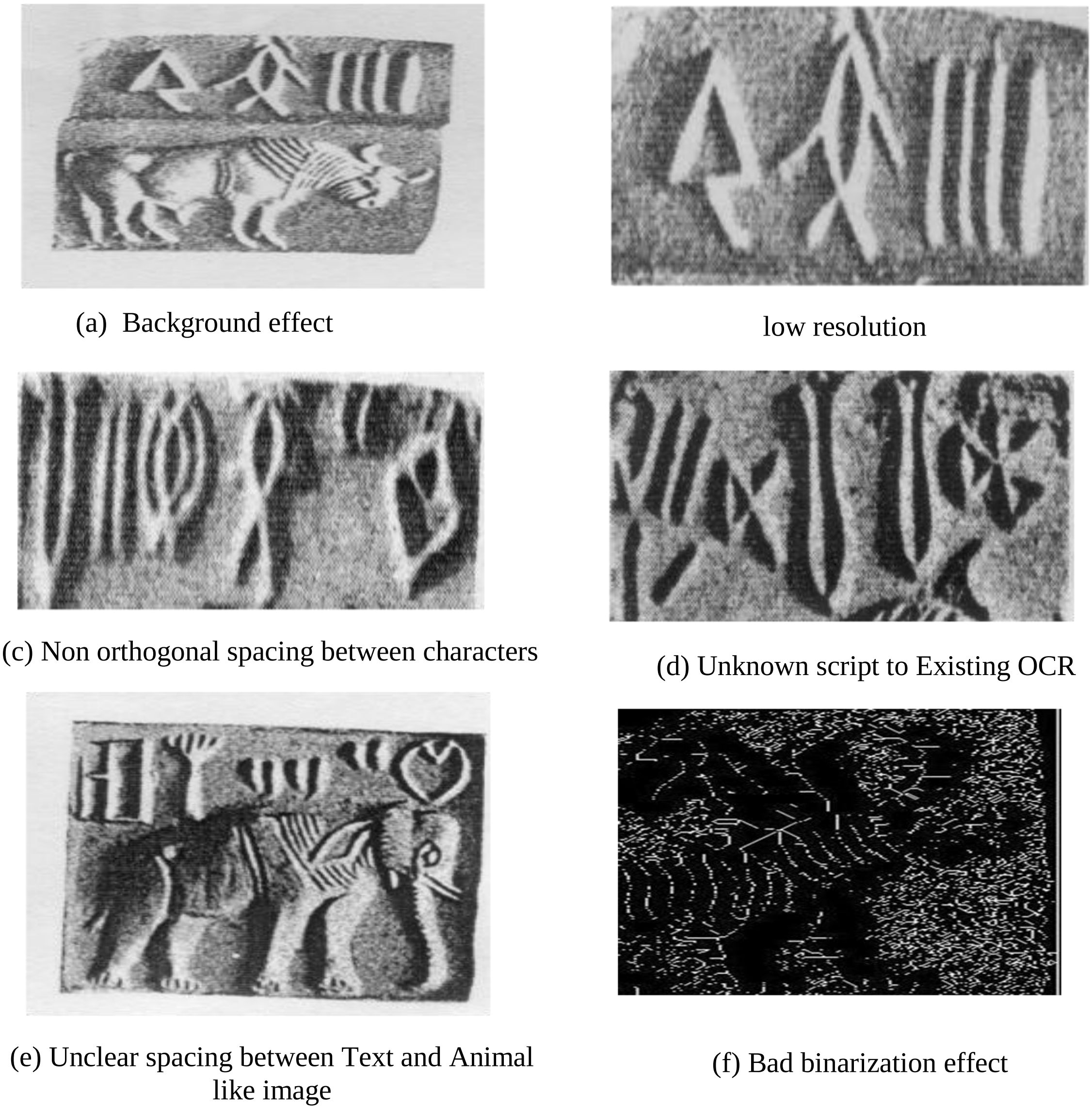
Usually document consists of texts along with animal like picture as shown in Fig. 1.5 and Fig. 1.6. The presence of animal like picture makes problem more challenging.

Since Indus documents engraved on hard material, such as rock, wall, palm leaves etc, the documents have complex background with poor quality and irregularity. This makes more challenging to segment text from background.

Sample images shown in Fig. 1.6 show that the images has poor quality due to aging and material that used for writing. Therefore, it is necessary to enhance the quality of the image. During that period, there are no pen and ink to write, people use stick to write on rock, palm leaves and wall. As a results, text appeared in Indus document image have no regular shapes like alphabets of modern language. Therefore, developing OCR for recognizing such text is hard.

Since text appeared like handwritten by different people, one can expect irregular spacing between the words and characters. Therefore, it is hard to segment text and character for recognition.

Text appeared has no boundary of style and shapes, which results in more cursive nature of character components. This is another challenging for developing OCR for recognition.



(a) Background effect

low resolution

(c) Non orthogonal spacing between characters

(d) Unknown script to Existing OCR

(e) Unclear spacing between Text and Animal like image

(f) Bad binarization effect

Fig. 1.6. Sample images showing problems of Indus scripts

1.4 Proposed Plan

As noted challenges in previous section, it is required novel idea for understanding Indus documents. In this thesis, we propose the following plan to achieve it. In general, Indus documents mixed with other documents, such as English, Tamil, Kannada, Telugu, Malayalam in Archeology department, there is a need for developing Indus script identification such that the developed system can automatically choose Indus document to recognize. In this work, we focus on South

Indian scripts because the Archeology department where we collect Indus document is located in South part of India. It is seen that texts in Indus document has animal like image called as non text and also it is highly cursive compared to other language scripts. This observation leads to propose identification method.

It is noted from the above discussion that Indus documents have poor quality. To enhance the quality of the image such that subsequent steps perform well, we propose a new idea which works based on fact that pixel which represent text have high contrast and pixels which represent background have low contrast.

Since Identified Indus documents consists of text and non-text part, in order to recognize text in Indus document image, text lines are to be segmented from the non-text. The previous step enhances text information. The same idea has been explored to segment text line from non-text which considers regular spacing and high contrast of text pixels.

After segmenting text lines, to recognize the character of Indus document, we propose an idea for segmenting character components based on watershed model which explore spacing between the characters. The idea is that where there is a space between the character components, there will be more collection of water if touching exists between two character components due to complex background else there will water flow.

For recognizing segmented characters, we propose a new descriptor called Histogram Oriented Tangents (HOT). As inspired by the success of Histogram Oriented Gradients (HOG) for object recognition, we propose HOT for recognizing Indus character components which is invariant to font, font size and writing variations. The flow of proposed plan are shown in the form of block diagram as shown in Fig. 1.7 where we can see steps of the proposed plan to recognize Indus character components.

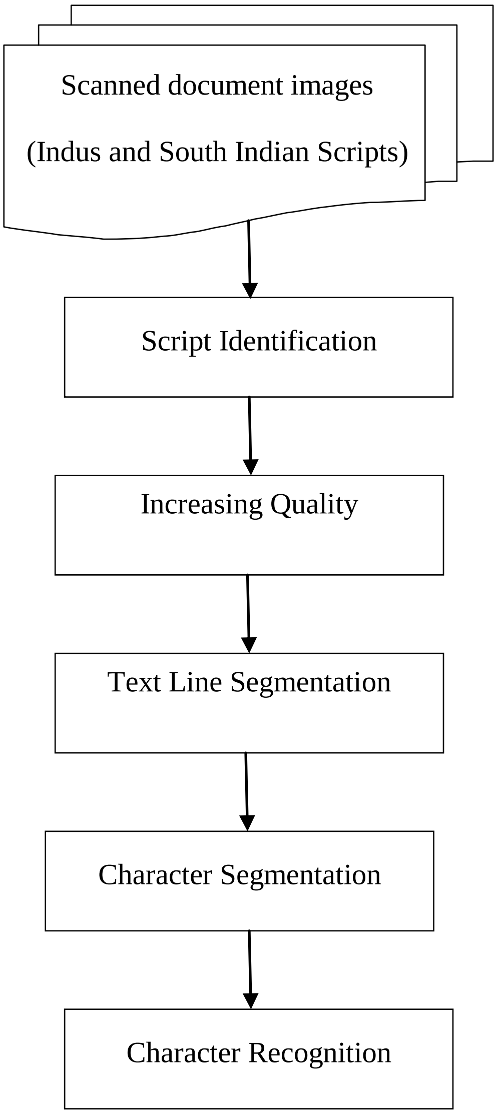


Fig. 1.7. Flowchart of proposed method in Indus document understanding

1.5 Contributions of the Thesis

List of contributions of the proposed thesis work are stated below.

- Developing a new idea based on cursiveness of Indus document and other documents for Indus document identification from south Indian scripts, namely, Kannada, Telugu, Tamil, Malayalam and English.

- To improve the quality of the Indus document images, the new idea has been introduced based on the combination of Laplacian and Sobel operation.

- Based on contrast information and spatial relationship between the pixels of text information, we propose a new idea for segmenting text lines from non-text information.

- Watershed model is explored for character segmentation from segmented text lines, which works well for touching between character components due to background complexity and non-uniform spacing between the character components.

- A new descriptor called Histogram Oriented Tangents (HOT) as inspired by Histogram Oriented Gradients (HOG) for recognizing character components which works well for different font, font size, writing style.

1.6 Organization of thesis

To present our findings, the proposed framework of the thesis is organized as follows.

Chapter 1 describes history of different types of documents, Outline of the thesis, Framework of the proposed methods, Motivation, Contributions of the research.

Chapter 2 describes various works that we encounter during our research. We exploit the detail survey of work related to Scanned document; Camera based images, Plain text images, Historical images and degraded images. We also discuss methods related to Identification of scripts, Segmentation of Text and Non text, Segmentation of Characters and Recognition of Characters.

Chapter 3 focus on methods developed for identification of Indus scripts from other scripts. First method focus in checking the position of Centroid to determine straightness and cursiveness using Morphological operations and region growing. To overcome drawbacks of first method, Skewness Approach and nearest neighbor approach is used for Identification of Indus scripts.

Chapter 4 describes enhancement method to enhance Text and Non text components from background. For the enhanced Indus document image, a new approach of boundary growing is proposed to segment Text lines from Non text.

Chapter 5 describes a method for segmentation of Indus Characters from text lines. For the purpose of character segmentation, we explore Watershed model for segmentation of characters from the segmented text lines.

In Chapter 6, a new hybrid approach for Indus text recognition is described. This method is applied to the resulting image of Chapter 5. Due to unpredictable nature of Indus texts a new histogram oriented tangent approach is proposed for recognition of Indus characters. Slope of boundary pixels and runs of pixels encountered in different directions used as feature for recognition.

Finally, Chapter 7 gives summary of the entire research work and scope for future work in each Chapter. The Chapter describes drawbacks and failure in different cases in each method and suggestions for further openings of research in different aspects. Limitations of the research work described in this thesis are also presented.

Chapter 2

Literature review

In previous chapter, we discuss about importance, usefulness of the Indus document recognition along with the list of challenges to achieve objective of the thesis. In addition, we also present our plan which consists of script identification, enhancement, text line segmentation, character segmentation and character recognition to recognize Indus document. In this chapter, we review existing methods of respective steps of our plan and other documents as well which include methods of scanned document images, camera based images, degraded document images, historical document images and Indus document images.

2.1 Methods of Scanned Document Image Analysis

According to literature review on document image analysis, it can be classified the methods of scanned document images into the methods of binary document images and the methods of plain document images. This section focus on review of both the methods to understand the state of the art in understanding the conventional methods for document image analysis.

2.1.1 Methods of Binary Document Images

In general, most of the time document image consists of homogeneous background with plain text. Therefore, the methods proposed to use binary image of the input image to reduce the complexity and number of computations to process the document image. In other words, these methods consider binary image as an input for document image analysis.

Robust Watermarking scheme for protection of document image contents is proposed by Chetan et al., 2015. Document image is divided into empty and non-empty segments depending on the absence or presence of the information. Watermarking is applied for non-empty segments and thus the amount of embedding capacity is reduced. A binary watermark logo is compressed using binary block coding technique of appropriate block-size. A level-2 integer wavelet transformation is applied on the non-empty segment of the source document image. The extracted and embedded watermarks are compared and authentication decision is taken based on majority voting technique. Based on the quantization step size, size of the logo and the level of wavelet transform, the watermarks are extracted without accessing the original image.

A skew-estimation method using straight lines in document images is presented by Hyung et al., 2016. Method exploit the properties of text, and formulate the skew-estimation problem as an estimation task using straight lines in images and focus on robust and accurate line detection. Block-based edge detector followed by a progressive line detector is used to check boundaries of figures/tables, vertical/horizontal separators, and boundaries of text blocks.

Information spotting in document image is presented by Pau et al., 2016. Method is Graph-based a binary embedding is defined as hashing keys for graph nodes. Labeled graphs, graph nodes are complemented with vectors of attributes representing their local context. Then, each attribute vector is converted to a binary code applying a binary-valued hash function. Therefore, graph retrieval is formulated in terms of finding target graphs in the database whose nodes have a small Hamming distance from the query nodes, easily computed with bitwise logical operators. Method is used in real scenarios such as handwritten word spotting in images of historical documents or symbol spotting in architectural floor plans.

Table detection in the document image is proposed by Tuan et al., 2016.This work propose a method for detecting table regions by using a new shape which is called Random Rotation Bounding Box. This shape is used for illustration and description of the table regions. Based on it, system performs the following three fundamental steps to detect the table zones: classification of the text and non-text elements in the document image, detection of the ruling-line tables, and identification of the non-ruling-line tables. The method also works when tables are skewed.

It is observed from the review on binary document images that despite the methods save number of computations, sometimes, the methods lose accuracy for the documents which have low quality, degradations and low contrast because of loss of information during binarization process. This is the major issue with the methods of binary document images. Therefore, these methods cannot be used for Indus document image analysis because Indus document suffer from multiple adverse factors as discussed in the chapter-1.

2.1.2 Methods of Plain Document Images

In order to avoid loss of information during binarization process, the methods prefer to use gray images without binarization for document image processing. In this way, the methods proposed to overcome the problems of the methods of binary document images.

Underline detection and removal from scanned documents to make text legible and improve OCR efficiency. Also the method concentrates on removal of several annotations. This method is proposed by Tadbeer et al., 2016 uses connected component and edge analysis using geometric rules. FCM clustering and ANN training and testing used to separate the unwanted region that is the annotated area from the scanned text which is required as final output.

Techniques to identify printed and handwritten text in scanned documents proposed by Tan et al., 2015 this paper address, the question of how to discriminate between each type of writing on registration forms. Registration-form documents consist of various type zones, such as printed text, handwriting, table, image, noise, etc., so, for segmentation the various zones called multiscale region projection used to identify printed text and handwriting. A new set of projection features extracted from each zone is also proposed. The classification rules are mining and are used to discern printed text and table lines from handwritten text.

Optical Character Recognition of Arabic handwritten characters proposed by Rana et al., 2015 extracts certain properties of each of the letters from scanned documents. The properties also called features are extracted from the plain document image to uniquely define character by the means of ANN. The method attempts for conversion of scanned image to machine encoded text that is used to implement OCR system. Method is designed for Arabic characters.

Noise removal from scanned images was done by Arnon et al., 2013 by analyzing image histogram or projection profile of the document image. Method detects marginal noises by finding the highest peaks of the projection profile in the corresponding regions. The marginal noises usually appear as dark regions around the margin of the scanned documents. Different steps involved in the method are: Creating image profile; Smoothing the profile by using moving average technique; Dividing the projection profile of the document image into three regions: left, middle, and right; Detecting marginal noises by finding the highest peaks of the projection profile in the corresponding regions; Removing the detected marginal noises by replacing the regions of marginal noises with white pixels.

Though the methods use gray information instead of binary information for document image processing, the methods assume the documents must have high contrast and plain background for achieving better results. This is not true for the document like Indus where one can expect multiple causes. In addition, since devices, such as scanner has inherent limitations, sometimes, devices introduce unpredictable noise, skew and distortions during scanning. Therefore, these methods may not work well for complex document images.

2.2 Methods of Camera Document Image Analysis

Since devices like scanner induce noise during scanning the images, consume more time and not user friendly compare to camera device, the images care captured by camera devices for document image processing. The limitations of scanner degrade the performance of the methods for degraded and historical document images because these documents suffer from low quality, noise, low contrast etc. Therefore, this section focus on the methods for degraded document images and historical document images.

2.2.1 Methods of Degraded Document Image Analysis

Documents are available in the form of paper. Degradation of documents can be due to folding, repeated touching, tearing, ink spell etc. We review some of the adaptive methods that work on degraded document image analysis. These images when captured by Camera are degraded by uneven light condition.

A Binirization method to overcome the problem is proposed by Jung et al., 2009. A descriptor that captures the regional properties around a given pixel is first defined for this purpose. For each pixel, the descriptor is defined as a vector composed of filter responses with varying length. This descriptor is shown to give highly discriminating pattern with respect to the background region, text region, and near text region. The misclassified pixels are relabeled using an energy optimization method, using the graph cut method. The proposed descriptor is also used for the skew detection, and thus correcting the skewed documents.

Scale and Rotation Invariant Features for Camera-Based Document image retrieval is proposed by Dang et al., 2015. The method deal with the occurrence of errors due to feature point extraction in Camera captured documents. Features are computed based on Geometrical constraints between pairs of nearest points around a keypoint.The method is based on Locally Likely Arrangement Hashing (LLAH), which has been widely used and accepted as an efficient real-time camera-based document image retrieval method .

Some of the binarization techniques discussed by Jyoti et al., 2015 for the improvement of degraded document images. The methods discussed are useful for OCR and document image retrieval. The techniques discussed overcome the problem of camera shaking or object movement while taking images in camera. Global methods like Otsu thresholding method which analysis the distribution of the gray values for binirazation. Whereas Adaptive methods define local regions R(x, y) and calculate a separate threshold value T(x, y) for each region.

The above methods are good for the images which have degradations with homogeneous background. Most of the methods focus on English text and recent document images but not historical document images where one cannot expect fixed fonts, font size, and shape of the character components. Therefore, these methods may not work for the documents which have large variations in text embedded in complex background.

2.2.2 Methods of Historical Document Image Analysis

Those documents may contain images of various artifacts, Un-Uniformly page coloring due to repeated handling, Paper quality, Dust, Environment conditions , breakages etc.Scripts may also ornamental with variable sizes. Hence, Historical document images captured in Camera will be different than degraded and other document images.

Enhancement of old images and documents by digital image processing techniques were proposed by, Preeti et al., 2015. Binarization techniques applied to remove the noise and improve the quality of the documents. Specialized processing is required to these document images for removing background noise in order to become more legible. A hybrid binarization approach is proposed in this paper for improving the quality for the old documents. Combination of global and local thresholding techniques are used for the same. Initially, a technique named global thresholding is applied to the whole image. The image area that still has background noise is detected and the technique is again re-applied to each area separately.

Enhancement of Historical Document Images by Combining Global and Local Binarization Technique is proposed by, Zemouri et al., 2014. The proposed hybrid approach includes both global and local thresholding techniques to deal with noisy historical documents, where firstly a global thresholding T is applied to whole document. The proposed method has been evaluated on two different datasets. First, an Arabic historical document dataset supplied by the National Library of Algeria is used based on word spotting system. Second, a Handwritten Document Image Binarization Contest dataset.

An Adaptive Multilayer-Information Binarization Technique for restoration of Degraded Historical Thai Document Image is proposed by Krisda et al., 2014. Method consists of five stages including noise elimination by applying the existing filtering technique, majority pixel analysis for extraction of foreground areas, degradation of the background layer estimation based on gray value, thresholding and vicinity analysis.

It is noted from the above discussion that the methods solves the problems of the document affected by degradations. However, the performance of the methods is good when the document images affected by single adverse factor. If the document image affected by multiple adverse factors, such as low quality, complex background, complex foreground etc, the methods report low results. Therefore, the methods may not be suitable for the document images like Indus where we can expect both complex background and foreground due to carving on rock, wall and palm leaves with sticks. In addition, most of the images contain animal like picture along with the text information. This makes more problems for the developed methods.

2.2.3 Methods of Indus Document Image Analysis

According to literature, there are not many methods for Indus document image processing but there are methods which work on documents which are similar to Indus document, such as epigraphically document images. Therefore, in this section, we review the methods on both epigraphically document images and Indus document images as well.

Rajesh P.N.Rao proposes the probabilistic analysis of Undeciphered Indus script. He estimates using Markov and N-gram Models the transition probabilities P (si |sj) that sign i follows sign j. The obvious way of estimating P (si |sj) is to count the number of times sign i follow sign j, an approach equivalent to maximum likelihood estimation. The presence of statistically significant clusters of symbols with positional preferences suggests that there is sequential order in the Indus script. Characterization of Indus script that discuss entropy of the script is published by Rajesh et al., 2010

Computational techniques for inferring the syntax of undeciphered scripts is proposed by Nisha et al., 2014. Paper discuss on analysis of syntactic patterns using Cumulative frequency distribution of text Enders and text beginners. They used bigram or a first-order Markov model model of the Indus script for restoring signs in illegible Indus texts**.** The range of correlation is restricted to the nearest neighbor.

Segmentation of Indus texts is proposed by Nisha et al., 2008 based on the structural analysis of Indus Texts. . Paper gives comprehensive approach to segment the Indus texts using statistically significant signs and their combinations in addition to all the texts of length 2, 3 and 4 signs. By this method and hence it can be suggested that the texts of 5 or more signs can actually be seen as permutations of other frequent sign-combinations or smaller texts (of length 2, 3 or 4 signs).

The above methods are limited to particular document images but not actual Indus document images which affected by multiple causes. Most of the methods follow the conventional way to attack the problem of Indus document images. This may not be robust approach to achieve better results for Indus document images. Therefore, Indus document image recognition is remains open issue in the field of document image analysis.

2.3 Methods of Script Identification

After concluding, there hardly methods for Indus document image processing, this section focus on identifying Indus script from other south Indian scripts and English script to make automatic system for recognizing Indus document. Therefore, we review methods of Latin script, non-Latin scripts and Indus script identification.

2.3.1 Methods of Latin Script Identification

The Latin alphabets are derived from the Etruscan and Greek alphabets and contributed many words to some of the languages like English, French, German , Italian , Spanish etc in western world. In this section, we review some of the methods for identification of Latin scripts. It is found that Latin scripts are a set of graphical signs. Latin script is used as the standard method of writing in most Western and Central European languages, as well as many languages from other parts of the world.

Identification of Latin-based Languages through Character Stroke Categorization proposed by Lu et al., 2007. The proposed technique detects languages through the word shape coding, which converts each word image into a word shape code and accordingly transforms each document image into an electronic document vector. For each Latin-based language under study, a language template is first constructed through a corpus-based learning process. The underlying language of the query document is then determined based on the similarity between the query document vector and multiple constructed language templates. Detection of words and text lines through the projection profile analysis and the illustration of the top line, bottom line, x line, base line, and middle line of text.

Latin script identification proposed by Lu et al., 2006 for degraded and distorted document images through document vectorization, which transforms each document image into an electronic document vector that characterizes the shape and frequency of the contained character and word images. Identification is based on the density and distribution of vertical runs between character strokes and a vertical scan line. Latin-based languages are then differentiated using a set of word shape codes constructed using horizontal word runs and character extremum points.

The above methods use language specific features for script processing. As a result, we cannot use the same methods directly for other scripts like non-Latin scripts.

2.3.2 Methods of Non-Latin Script Identification

We can see from survey Scripts like Chinese, Arabic, Brahmi, Bengali are called Non-Latin Scripts which have their own existence and are not evolved from Latin. It is essential to look back the identification methods existing before solving the problem of our research. Here is a highlight of the methods to identify Non Latin Scripts that considers image captured in various ways.

Another Ancient Non Latin script that to be known is Brahmi Script evolved earlier that gives rise to other scripts like Devanagari, Oriya, Telugu and Kannada. Script identification methods for Tri lingual image document is presented by Anil et al., 2014. The main objective of this system is to identify the specific script and feed them into their specified Optical Character Recognition (OCR) system. OCR is the system which converts the image document into editable text document. Script identification of written text in the domain of Indian script based languages is a well-studied research field. In this paper a technique of script Identification is described to discriminate three major south Indian scripts: Oriya, Telugu and Kannada. The approach is based on the analysis of horizontal projection and vertical projection profile.

Identification of Non Latin script such as Chinese and others Non Latin Scripts is proposed by Shiva kumara et al., 2011. Method proposed for video script identification. They extract upper and lower extreme points for each connected component of canny edges of text lines. The extracted points are connected to study the behavior of upper and lower lines. The direction of each 10-pixel segment of the lines is determined using PCA. The average angle of the segments of the upper and lower lines is computed to study the smoothness and cursiveness of the lines. In addition, to discriminate the scripts accurately, the method divides a text line into five equal zones horizontally to study the smoothness and cursiveness of the upper and lower lines of each zone.

Arabic Script identification of documents based on letter frequency using a back propagation neural network is proposed by Ali et al., 2005. Paper analyzes the feasibility of using a windowing algorithm in order to find the best method in selecting the features of Arabic script documents language identification using back propagation neural networks and used the datasets belonging to Arabic, Persian, Urdu and Pashto language documents which are all Arabic script languages.

Latin and non Latin script identification methods using Steerable Pyramid Features is proposed by Mohamed et al., 2009. The features extracted from pyramid sub bands serve to classify the scripts .Similarly, script identification methods for Latin and Non Latin is proposed by Lijun et al., 2006. Paper discuss on the methods based on Connected Component profiles. It shows the identification methods and distinguished features for Bangla and English scripts. To extract features from the text block, the set of connected components in the destination address block image is calculated first. Since Bangla text is cursive i.e. characters are connected within each word, a connected component in Bangla text image may correspond to a word. In contrast, English characters are isolated unless conditions due to low print quality or poor scanning. Thus, a connected component is generally related to a character in printed English text block.

However, the above methods are sensitive to degradations as these methods use connected component analysis based features.

2.3.3 Methods of Indus Script Identification

Since there are no direct approaches for Indus script Identification, we review the methods on scripts which are similar to Indus, such as historical documents.

Probabilistic models are proposed by Rajesh et al., 2009 to analyze the structure of the Indus script. The goal is to reveal, through probabilistic analysis, syntactic patterns that could point the way to eventual Decipherment. Simple Markov chain model used to capture sequential dependencies between signs in the Indus script. The trained model allows new sample texts to be generated, revealing recurring patterns of signs that could potentially form functional subunits of a possible underlying language. The model also provides a quantitative way of testing whether a particular string belongs to the putative language as captured by the Markov model. The likelihood of a particular sequence of Indus signs with respect to the learned Markov model is discussed to know the sign sequence belongs to the putative language encoded by the Markov model.

Sequences of Indus signs are analyzed by Nisha et al., 2011 which demonstrate presence of a rich syntax and logic in its structure. Here structural design of individual signs of the Indus script. Our study is based on the sign list given in the concordance of Mahadevan (1977) which consists of 417 distinct signs. Analysis of the structure of all signs in the sign list of Indus script and visually identifies three types of design elements of Indus signs namely basic signs, provisional basic signs and modifiers. These elements combine in a variety of ways to generate the entire set of Indus signs. By comparing the environment of compound signs with all possible sequences of constituent basic signs, we show that sign compounding (ligaturing) and sign modification seem to change the meaning or add value to basic signs rather than save writing space. The study aims to provide an understanding of the general makeup and mechanics of design of Indus signs.

Overall, most of the methods proposed for identification of different scripts use connected component based features. These features may not work well for Indus documents. In addition, none of the methods proposed specifically for Indus script identification from especially from south Indian scripts where we can expect common features shapes.

2.4Methods of Text Line Segmentation

To recognize the text in the document, one should segment text lines from the document image. In this section, we review the methods on text line segmentation from plain document images, degraded images, historical document images and Indus document images.

2.4.1 Methods of Plain Background Document Images

In this case, the methods accept either binary image or gray image with plain background images for text line segmentation.

Christopher et al., 2015 describes a novel technique for text line extraction based upon seam carving. Methods, direct area detection and information masking, are described. The seam carving approach attempts to find whitespace seams directly. Energies of each pixel found. Using results of an energy function, seams must next be calculated for the image. Seam value is calculated for each pixel by adding the minimum of the left three pixels in the current pixel's 8-connected neighborhood (left above, left, left below) to the current pixel's energy value.

Text-line extractions from handwritten documents have been proposed by Jewoong et al., 2014. Method is based on connected components (CCs), analyze strokes and partition under-segmented CCs into normalized ones. Due to this normalization, the proposed method is able to estimate the cost function is built whose minimization yields text-lines.

Text Line Detection for Heterogeneous Documents proposed by Markus et al., 2013. Segmentation of text lines based on the distances among the bounding boxes of components in an image. Foreground elements are grouped to words using Local Projection Profiles (LPP). The text line detection presented is a bottom-up approach which utilizes profile boxes of words as basic entity. These words are merged according to their minimal distance. Then, line rectangles are calculated using PCA for robust line orientation estimation.

Since the methods assume plain background and binary images, the methods may not work well for complex background and non-uniform spacing between text lines images.

2.4.2 Methods of Degraded Document Images

In this case, the methods considers gray image with some degradations as input for text line segmentation.

Text-Line Detection in Camera-Captured Document Images Using the State Estimation of Connected Components was proposed by Hyung et al., 2016. Method is developed by incorporating state estimation (an extension of scale selection) into a connected component (CC)-based framework. To be precise, CCs are extracted with the maximally stable region algorithm and estimate the scales and orientations of CCs from their projection profiles. Since this state estimation facilitates a merging process (bottom–up clustering) and provides a stopping criterion, method is able to handle arbitrarily oriented text-lines and works robustly for a range of scales. Finally, a text-line/non-text-line classifier is trained and non-text candidates (e.g., background clutters) are filtered out with the classifier.

Method for noisy and complex background known as Global Optimal Text Line Extraction based on K Shortest Paths algorithm is proposed by Liuan et al., 2015. Method is based on K-shortest paths global optimization in images. Firstly, the candidate connected components are extracted by reformulating it as Maximal Stable Extreme Region (MSER) results in images. Then, the directed graph is built upon connected component nodes with edges comprising of unary and pair wise cost function. Finally, the text line extraction problem is solved using the k-shortest paths optimization algorithm by taking particular structure of the directed graph.

Text line extraction method in Scanned Document Gray Scale images was proposed by Raid et al., 2014.The method is proposed directly on gray scale document images. In this paper a new approach for text line segmentation that works directly on gray-scale document images. Distance transform is applied directly on the gray-scale images, which is used to compute two types of seams: Medial Seams and Separating Seams. A Medial Seam is a chain of pixels that crosses the text area of a text line and a Separating Seam is a path that passes between two consecutive rows. The Medial seam determines a text line and the separating seams define the upper and lower boundaries of the text line. The Medial and Separating Seams propagate according to energy maps, which are defined based on the constructed distance transform. Scanned Gray scale images have noise. The distance transform of noisy document images may include small fluctuation that influences Seam generation. To overcome this limitation, Gaussian filters to smooth the image before generating the distance transform.

Text Line Detection in Corrupted and Damaged Historical Manuscripts is proposed by Rabaev et al.2013. Method grouped text lines by analyzing the evolution maps of connected components. A sweep line moved from left to right is further used to check whether elements lie in the same line. However, the method can only detect lines of equal-size texts which are chosen in their dataset. Method is found to be powerful to detected characters in torn and damaged Manuscripts.

The method extracts features which are invariant to degradations for text line segmentation. However, the method requires homogenous background images for achieving better results.

2.4.3 Methods of Historical Document Images

In this case, the method considers the documents which suffer from low quality and degradations as input for text line segmentation.

Vesselness for text detection in historical document images is proposed by Simon et al., 2016.This paper describes a method to detect text in images, particularly in historical document images. For a robust detection, vesselness filter is used on the grayscale document image for text detection, which gives a probability for each pixel being text. At the locations segmented by this filter, SIFT key points are detected which are spatially clustered. Overlapping windows from these clusters are subsequently VLAD encoded and classified in text and non-text. ROOTSIFT key points are extracted from the text mask, and key points are clustered by their spatial location. Each of the created clusters is then taken as input to a sliding window approach, generating feature descriptors at the respective key point locations. These feature descriptors are further aggregated and encoded using VLAD and classified as either text or non-text by a linear classifier.

Global method for automatic text line extraction is presented by Raid et al., 2014. The proposed approach computes an energy map of a text image and determines the seams that pass across and between text lines. Two different algorithms proposed, one for binary images and the other for grayscale images. The first algorithm extracts the components along text lines. The second algorithm computes the distance transform directly from the grayscale images and generates two types of seams*:* medial seams and separating seams. The medial seams determine the text lines and the separating seams define the upper and lower boundaries of these text lines.

Binarization-free clustering method proposed by Angelika et al., 2013 segment curved text lines in Historical documents. Complex layouts on the one hand, such as curved and touching text lines and binarization problems on the other hand, caused by ornaments, wrinkles, stains, holes, etc. In this paper, we propose a binarization-free clustering method for text line segmentation that is not only able to cope with touching text lines, but also with complex baseline curvature. Avoiding the assumption of straight baselines, small interest point clusters are grouped into text lines based on their local orientation. Experiments conducted on artificially distorted images of the Saint Gall database. Document image is transformed as a set of interest points describing parts of characters such as Junctions, stroke endings, and circular structures.

Another, Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering proposed by Angelika et al., 2012 . Interest points representing parts of characters are extracted from gray-scale images. Next, word clusters are identified in high-density regions and touching components such as ascenders and descenders are separated using seam carving. Finally, text lines are generated by concatenating neighboring word clusters, where neighborhood is defined by the prevailing orientation of the words in the document.

The methods work well for poor quality images and limited degradations. However, when the documents contain irregular spacing between text lines with complex background, the methods fails to perform well.

2.4.4 Methods of Indus Document Images

In this case, the methods consider the documents like Indus for text line segmentation.

It is found some of the Indus texts are long; a new approach by Gift et al., 1988 was presented using a dynamic programming model to segment short texts from the long texts to show that long inscriptions are the combination of more than one text. Demonstrates how a dynamic programming algorithm can be developed to segment unusually long written inscriptions from the Indus Valley Civilization. Explains the problem of segmentation, discusses the dynamic programming algorithm used, and includes tables which illustrate the segmentation of the inscriptions.

Over all, it is concluded that the methods developed for other document images may not work well for the Indus document images. The main reason is that Indus document contains usually animal like picture along with text information. This makes confusion about the spacing between the text lines.

2.5 Methods of Character Segmentation

To recognize characters from the text line, one has to segment characters from text lines. In this section, we consider text lines as input for character segmentation. Therefore, we review the methods for character segmentation from plain background images, degraded document images, historical document images and Indus document images.

2.5.1 Methods of Plain Background Text Line Images

The methods consider text line images with plain background for segmenting character components.

Touching character segmentation of Devanagari script is proposed by Subith et al., 2016. Method proposed for Devanagari script, a two dimensional form of symbol. Segmentation of offline handwritten document especially touching character segmentation is discussed in this paper. Paper also discusses problem and processes in touching character segmentation.

Bangla Handwritten Character Segmentation Using Structural Features proposed by Tapan et al., 2016 to Segment Bangla Handwritten Word Images into meaningful individual Symbols or Pseudo-characters. Segmentation algorithm is Two-class Supervised Classification problem. The method employs an SVM classifier to select the segmentation points on the word image on the basis of various structural features.

Character Segmentation of Hindi Unconstrained Handwritten Words is proposed by Soumen et al., 2016. Method handles inherent variability during segmentation of Cursive Hindi characters. Segmentation is performed on the basis of some Structural Patterns observed in the writing style of this language. The proposed method can cope with high variations in writing style and Skewed Header lines as input.

Similarly, segmentation of offline characters in Cursive Uyghur handwriting words is proposed by Mayire et al., 2012. Input for this method is Uyghur characters in cursive form which is the combination of 2 or 4 shapes. Method removes delayed strokes from the handwritten words and potential breakpoints are detected from concavities and ligatures by temporal and shape analysis of the stroke trajectory for segmentation point detection.

Segmentation of Touching Characters Using Contour Based Shape Decomposition is proposed by Le et al., 2012. The shape contour is linearized into edge lets and edge lets are merged into boundary fragments. The connection cost between boundary fragments is obtained by considering local smoothness, connection length and a stroke-level property called the Stroke Rate. Samples of connections among boundary fragments are randomly generated and the one with the minimum global cost is selected to produce the final segmentation of the shape.

Since the methods are developed for plain background and binary images, the methods are sensitive to degradations, noise and distortions.

2.5.2 Methods of Degraded Text Line Images

In this case, the methods consider gray text line images with degradations for character segmentation.

Mean-Based Thresholding Approach for Broken Character Segmentation from Printed Gujarati Documents is proposed by Riddhi et al., 2015. The method is proposed to segment broken character in machine printed Gujarati document image. This character degrades the image. Broken characters are generated due to Noise Scanning, Older Documents with Low-Quality Printing, and Thresholding error. It is necessary to identify and segment it properly. So this paper presents Mean-Based Thresholding technique for broken character segmentation from Printed Gujarati documents. Heuristic information is used to merge the identified broken characters. Methods merge Vertical Broken Gujarati characters as a single glyph from the document image.

A Robust Segmentation Technique for Line, Word and Character Extraction from Kannada Text in Low Resolution Display Board Images is proposed by Shanmukhappa et al., 2015.The proposed method uses Projection Profile features and on Pixel Distribution Statistics for Segmentation of Text Lines. The method also detects text lines containing consonant Modifiers and merges them with corresponding text lines, and efficiently separates Overlapped Text Lines as well. The Character Extraction process computes character boundaries using Vertical Profile features for extracting character images from every text line. The proposed method is tolerant to font variability, spacing variations between characters and words, absence of free segmentation path due to consonant and vowel modifiers, noise and other degradations.

Similarly, Segmentation of Broken and Isolated characters in Handwritten Gurumukhi Word using Neighboring pixel technique proposed by Akashdeep et al., 2015. Combination of two approaches like Horizontal Profile Projection and Vertical Profile Projection is used for segmentation. Neighboring Pixel algorithm is used to isolate broken and touching characters in Gurumukhi script.

Restoration and Segmentation of Highly Degraded Characters Using a Shape-Independent Level Set Approach and Multi-level Classifiers is proposed by, Reza et al., 2009. The method combines 2 approaches for segmentation. In first level, multi level classifiers consider stroke width information to locate candidate character pixels. In second level, a set of active contour scheme is used to identify boundary of character. Stroke cavity map and estimation of intensity are considered in the method for segmenting degraded characters from degraded background. Tests are conducted on ancient degraded Hebraic character images.

The methods are good for the images which have homogenous background images where we can see clear space between characters. However, this is not true for complex document images like Indus where we cannot expect regular spacing between characters.

2.5.3 Methods of Historical Text Line Images

In this case, the methods consider historical text line images for character segmentation. The documents usually suffer from poor quality due to aging.

To cope with the challenges in Historical Documents like Noise and Degradation, Overlapping Layouts, great variability of Page Layout, a method based on Learning Texture Features for Historical Document Image Enhancement and Segmentation is proposed by Maroua et al., 2015. The proposed method is based on using the Simple Linear Iterative Clustering (SLIC) super pixels, Gabor descriptors and Support Vector Machines (SVM). The proposed method provides interesting results on Historical Document Images having various Page Layouts and different Typographical and Graphical properties. Results tested on HBR2013 dataset.

Character segmentation for Ancient Telugu text documents is proposed by Venkata et al., 2015. By using analysis of Canonical Character Segmentation technique. A Hybrid model that entails Segmentation in noisy images followed by binarization is proposed to remove the noise in ancient images. In the first phase, segmentation technique for the ancient Telugu document image into meaningful units is proposed. Horizontal profile pattern is convolved with Gaussian kernel. The statistical properties of meaningful units are explored through an extensive analysis of the geometrical patterns of meaningful units in the First phase. In the second phase, noisy documents are cleaned with the help of Modified IGT algorithm and then segmented by using conventional profile mechanism. The performance of the present hybrid technique is proved by the results of higher efficiencies for the cleaned documents. The efficiency analysis of Segmentation carried out for the present Hybrid technique reveals a threshold number of Vowels (V), Consonants(C), and CV core characters to exhibit higher efficiencies.

Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques is proposed by, Sridevi et al., 2012. Ancient Tamil scripts documents consist of Vowels, Consonants and various Modifiers. Method uses Projection Profile and PSO algorithm for Line Segmentation. Combination of connected components along with Nearest Neighborhood methods are used to segment the Characters.

Segmentation of characters from Korean historical documents is proposed by, Min et al., 2004. In this paper, a method is proposed for segmentation and rejection methods for handling complex layouts. Proposed recognition-based segmentation method uses geometric feature and context information with Viterbi algorithm. Rejection method uses Mahalanobis distance and posterior probability for solving out-of-class problem.

Though the methods work well for poor quality and degraded images, the methods fail to give satisfactory results for the complex background images such as Indus document where animal like picture affect regular spacing between the characters.

2.5.4 Methods of Indus Text Line Images

In this case, the methods consider the document which is similar to Indus if available for character segmentation.

Segmentation of Indus texts using traditional method is proposed by Nisha et al., 2008. Segmentation is done using 4 methods. In first method two texts are compared which are identical by comparing few signs beginning or end. In second method using frequent combinations of signs like two-signs, three-signs etc. which can be treated as segments or identifiable units merely by their frequent rate of occurrence. In third method using sign-pair frequencies: The strongest and weakest junction points in a text based on the frequency of adjacent sign-pairs can be used for segmentation. Fourth method use Single Signs: Single signs falling in the categories of frequent beginners, frequent Enders, and frequent auxiliary Enders can be used to segment these texts.

Overall, the developed methods expect regular spacing between the character images and simple document images to achieve better results. Therefore, one can assert that these methods may not work well for the Indus document images because of irregular spacing between the characters and irregular shapes of the character symbols.

2.6 Methods of Character Recognition

After reviewing methods on character segmentation from text lines of different types of document images, this section focus on the methods for character recognition by considering segmented character as input for recognition. This also includes review on character recognition from plain/binary character images, degraded character images, historical character images and Indus character images.

2.6.1 Methods of Binary Character Images

The methods consider binarized character images for recognition as it is simple to develop a method.

Handwritten Devanagari Character Recognition using Local Binary Pattern is proposed by Prabhanjan et al., 2016. LBP operator is used for Circular Neighborhoods. The operator uses Neighborhood of different sizes. Sampling points is used as feature points on a Circle of radius R. Then characters are assigned labels based on feature extracted and the relationship among features. SVM is used for recognition.

Neural network based character recognition system proposed by Priyadarshni et al., 2016. In the proposed work a Character Recognition System to extract printed text from an image is developed using Kohenen self organizing maps (SOM) based retrieval system. SOM being an unsupervised method of training has a superior feature extracting property. Samples of same characters which are oriented at same angle but with different size, color and fonts are used. After calculation of certain Topological and Geometrical properties of a character it is classified and recognized. With self organizing map together with K means clustering algorithm.

A recognition method for Online Character is proposed by Minoru et al., 2014. Method focuses importance of Global features in Online Character Recognition. Global features represent the relationship between two temporally distant points in a Handwriting pattern. O (N^2) Global Features are extracted from a handwriting pattern with N points; selecting those that are truly discriminative is very important. AdaBoost is employed for feature selection. Experiments prove that many global features are discriminative and the combined use of Local and Global features can improve the recognition accuracy.

Since the methods depend on the performance of the binarization methods, these methods sometimes lose shape of the character information during binarization especially for low quality images.

2.6.2 Methods of Degraded Character Images

In this case, the methods consider gray character images with some degradation for character recognition.

Characters written on Palm leaves suffer from Degradation. This is because of Structure, Texture of Palm Leaves. Modeling of Palm Leaf Character Recognition System using Transform based techniques are proposed by, Narahari et al., 2016. Palm Leaf manuscripts which are very fragile and susceptible to damage caused by Insects, contain huge amount of information relating to Music, Astrology, Astronomy etc. The proposed work exploits a Special 3D feature (depth of indentation) which is proportional to the pressure applied by the scriber at that point. This 3D feature is obtained at each of the pixel point of a Telugu palm leaf character. In this work two dimensional Discrete wavelet transform (2-D DWT), two dimensional fast Fourier transform (2-D FFT) and two dimensional discrete cosine transform (2-D DCT) are used for feature extraction. The 3D feature along with the proposed two level transform based technique helps to obtain better recognition accuracy.

Similarly, a 3D approach to recognize Telugu palm leaf text is proposed by, Vijaya et al., 2016. Due to various factors like aging, insect bites, stains, etc., Palm leaves are easily susceptible to deterioration. Hence preserving and digitizing such fragile documents is highly essential. Traditional Scanning or Camera capturing of such documents suffer from multiple noise artifacts. A Depth Sensing approach is proposed to eliminate background noise for such manuscripts. The segmented characters extracted from Telugu palm scripts are further recognized using statistical approaches. The data points contain (X, Y, Z) co-ordinates for each segmented character. Selecting two co-ordinates at a time, patterns are generated in XY, YZ and XZ planes of projection.

To tackle the problem of blurness in Camera-Captured documents and to build High Performance OCR for Blurred Documents with LSTM Networks was proposed by Fallak et al., 2016.LSTM networks have been applied directly to the gray-scale document images to avoid error-prone binarization of blurred documents. Experiments are conducted on publicly available SmartDoc-QA dataset that contains a wide variety of image blur degradations. To reduce the impact of different image degradations, the method performs few pre-processing steps: (i) to remove the perspective distortion, the image is first converted to a gray-scale image. Then canny edge detector to extract the boundaries of the document. In order to segment the document from the background, we applied morphological operations (erosion, dilation and hole filling) on the image. To remove the noise outside the boundary of the document we find the connected components and filter out the components that have an area less than a predefined threshold. Then, Harris corner detection to detect the corners and filter the detected corners in order to extract the four corner points of the document. Finally, geometric image transformation to warp the perspective using the four corner points. (ii) Run Length Smearing Algorithm is then used for extracting text lines from the document images. (iii) The text-line images extracted from RLSA are normalized to a fixed height of 48 pixels.

Since the methods assume background of character images is homogeneous, the methods obtain successfully character shapes during binarization. However, when the character images have complex background and irregular shapes, the methods perform badly.

2.6.3 Methods of Historical Character Images

In this case, the methods consider character images from historical document images where we can expect poor quality for character recognition.

CNN Based Transfer Learning for Historical Chinese Character Recognition was proposed by, Yejun et al., 2016. Historical Chinese character recognition has been suffering from the problem of lacking sufficient labeled training samples. A transfer learning method based on Convolution Neural Network (CNN) for historical Chinese character recognition is proposed in this paper. A CNN model L is trained by printed Chinese character samples in the source domain. The network structure and weights of model L are used to initialize another CNN model T, which is regarded as the feature extractor and classifier in the target domain. The model T is then fine-tuned by a few labeled historical or handwritten Chinese character samples, and used for final evaluation in the target domain. Several experiments regarding essential factors of the CNN Based Transfer Learning method are conducted, showing that the proposed method is effective.

Nom Historical document recognition system is being developed by Truyen et al., 2016. Method performs image Binarization, Character Segmentation, and Character Recognition. It incorporates two versions of Off-line Character Recognition: one for automatic recognition of Scanned and Segmented Character Patterns (7660 categories) and the other for user Handwritten input (32,695 categories). Both versions use the same recognition method, but they are trained using different sets of training patterns. Recursive *X–Y* cut and Voronoi diagrams are used for segmentation; *k–d* tree and generalized learning Vector Quantization are used for coarse classification; and the Modified Quadratic Discriminant function is used for Fine classification. The system provides an interface through which a user can check the results, change Binarization methods, Rectify Segmentation, and input correct character categories by hand. Evaluation done using a limited number of Nom historical documents after providing ground truths for them showed that the two stages of recognition along with user checking and correction improved the recognition results significantly.

An Optical Character Recognition System (QATIP) for Arabic Heritage Collections in Libraries is proposed by Felix et al., 2016. Novel approaches for language modeling and ligature modeling for continuous Arabic OCR. We test our QATIP system on an early print and a historical manuscript. Sophisticated text image normalization prior to feature extraction. First, the method estimates the baseline using the Hough-based approach which can handle straight or partially linear baselines. In case of curved baselines, we provide a method which fits a low order polynomial to the extremes in ascenders. Both methods result in horizontal and rectified baselines. Second, we correct the text slant using Hough space derivatives. Third, we normalize the image height while taking into account the position of the rectified baseline. Slant correction and size normalization were also described.

Since characters from historical documents images are complex compared the characters from other document images, the methods propose to use classifier for recognition. The methods extract feature which are invariant to degradations and distortions to some extent and then they train the classifier with large number of training samples for recognition. These methods are performing better than the conventional methods. However, the methods scope is limited to particular script due to classifier and predefined number of images for training. Since Indus character recognition does not have fixed number of shapes and characters in addition to complex background, poor quality, low contrast, blur, the methods may not be suitable for recognition.

In summary, the review on methods for script identification, text line segmentation, character segmentation from text lines and character recognition shows that there are no specific methods developed for handling Indus documents from script identification to recognition. In addition, none of the methods addressed the problem of Indus document recognition. The main reason is that usually Indus documents images affected by multiple diverse factors, such as low quality, low contrast, complex background due the presence of animal like picture, degradations due to rock, wall, palm leaves, aging due to old documents, variety of writing style due to use of stick instead of pen and Ink, irregular shapes of the character components and irregular spacing between the text lines and characters etc. Hence, developing OCR for recognizing Indus document requires new method at each stage from script identification to recognition. These challenges motivate us to consider it as a research issue for our thesis work. Thus the objective is to develop new methods at each stage such that it results in OCR for Indus documents.

Chapter 3

Indus script identification

3.1 Background

The literature review presented in chapter 2 shows that historical-degraded document understanding is challenging especially Indus documents. It is true that archelogy department usually preserve such documents. However, it is noted that the Indus documents are dumped with other documents, such as English, Kannada, Tamil, Telugu and Malayalam etc. This observation leads to propose Indus script identification to separate from other documents. This is because the scope of this work is limited to Indus document recognition.

In this chapter, we propose a method for Indus script identification based on the fact that the components in the Indus document exhibits more cursiveness compared to the component in other documents. This is due to Indus documents usually contain animal like pictures along with text components and text is handwritten by sticks. In addition, since text is written over rocks, palm leaves, wall etc, nature of background also contributes to more cursiveness of the text components in Indus document.

Some parts of the material of this chapter appeared in the following research papers:

1. "An Integrated Method for Classification of Indus and English Document Images", International Conference on Emerging Trends in Electronics Computer Science and Technology (ICERECT-12), 21st December 2012.
2. "An Integrated Method for Classification of Indus and English Document Images",Lecture Notes in Electrical Engineering (LNEE), January 2014, Vol.248, pp.343-355.
3. "Skewness and Nearest Neighbour based Approach for Historical Document Classification", Proceedings of the 2013 International Conference on Communication Systems and Network Technologies(CSNT, IEEE Computer Society Washington, DC, USA), April 2013, pp. 602-606.
4. "A Robust Script Identification System for Historical Indian Document Images", Malaysian Journal of Computer Science(MJCS) , Vol. 28(4), 2015, pp 283-300 283.

3.2 Proposed Methodology

This section presents three different approaches for identification of scripts, namely, method based on Straightness and Cursiveness of the Components, the method based on Skew and Orientation of the Components and the method based on Spatial Relationship of Dominant Points of the Components.The block diagram of the method are shown in Fig. 3.1 for Indus script identification.

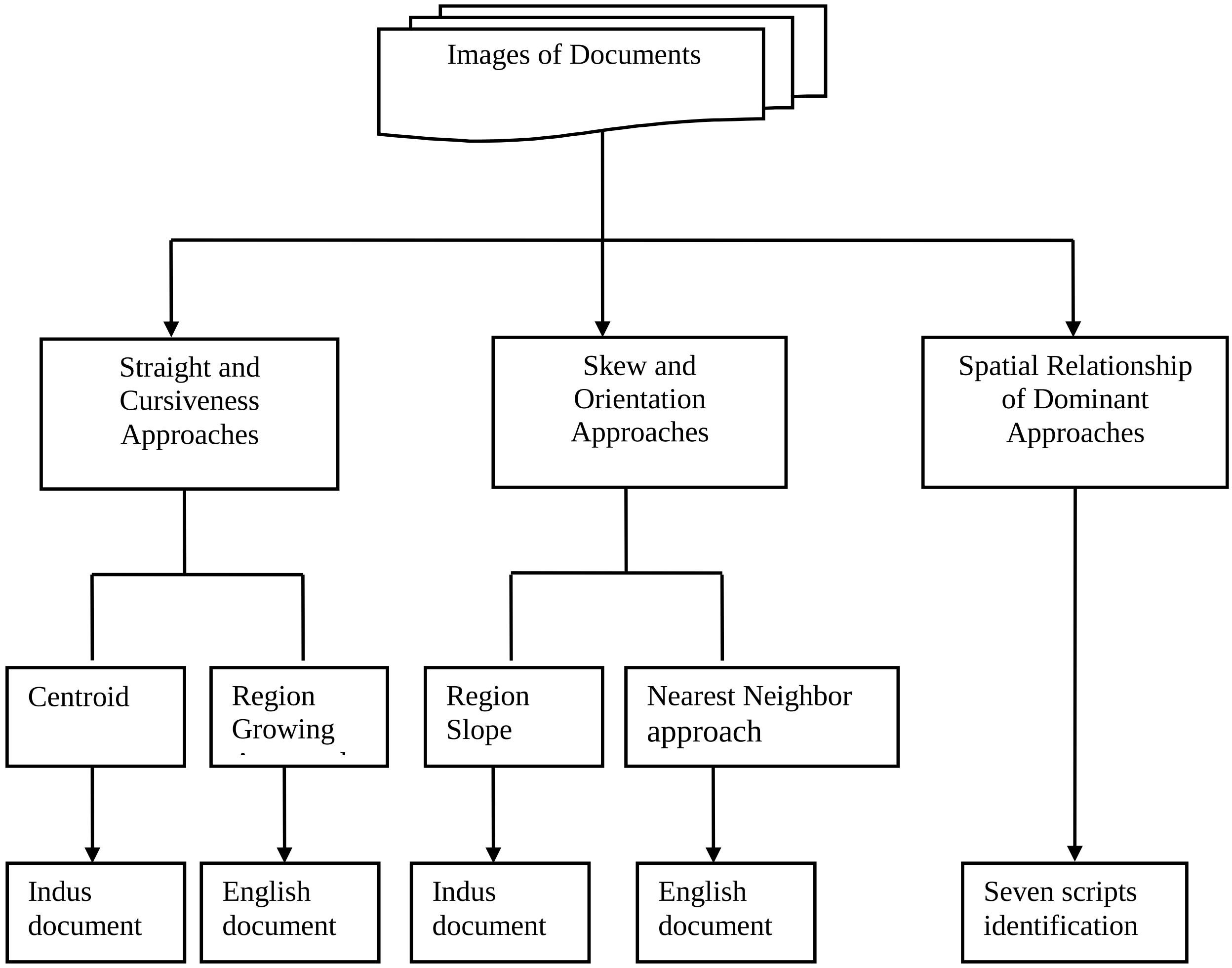

Images of Documents
Straight and Cursiveness Approaches
Skew and Orientation Approaches
Spatial Relationship of Dominant Approaches
Centroid
Region Growing
Region Slope
Nearest Neighbor approach
Indus document
English document
Indus document
English document
Seven scripts identification

Fig. 3.1. Proposed Methodologies

3.2.1 An Approach based on Straight and Cursiveness of the Components

This section proposes two ways to extracts Straight and Cursivenessfeatures for the components in the images of different scripts, namely, centroid based and region growing based ways for straight and cursiveness features. In this chapter, we consider south Indian documents, English and Indus document for script identification.

3.2.1.1 Centroid based Straightness and Cursiveness

It is fact that components in Indus documents have more cursiveness while other documents, such as south Indian documents images and English have more straightness compared to Indus document. Based on this observation, we propose a method to extract straightness and cursiveness for the components in the documents. The proposed method obtains Canny edge image for the input image and then check centroid whether it falls on the components or not. If it falls on the same component, it is considered as straight component else cursive components. The logic of method is shown in Fig 3.2.

For a given input image, the proposed method finds the Edges of images using Canny edge detector. To smooth the edges, the proposed method performs morphological operation such as, Dilation, merges disconnected components into single component as shown Fig.3.3(a)-Fig. 3(c). To study structure of the components, we perform thinning operation. This results in component with single pixel width as shown in Fig.3.3(d).Then the proposed method finds Centroid point (C) as shown in Fig.3.3(d) to study edge component is Straight or Cursive. Centroidis the pair of values that constitutes the mean valuesof Xco-ordinates and Y co-ordinates of edge components as defined in equation (3.1). Then the proposed method determines edge component is straight ifCentroid falls on edge components or within certain range. The Method computes the distance between edge component and Centroid using Equation (3.2). It considers the component as Cursive if Centroid is far from edge components. The threshold is determined as defined in equation (3.3) Furthermore, the method computes percentage of straight components and cursive components forwhole document. If the percentage of Cursiveness is more than the percentage of straight components, the document is classified as Indus document else south Indian documents. The percentage is computed as defined in equation (3.4) and equation (3.5). Note that the number of components in edge images is considered as actual number of components (total number of components) to calculate percentage.

To compute Centroid point, let $X = \{X1, X2, \dots, Xn\}\ and\ Y = \{Y1, Y2, \dots, Yn\}$ corresponds to X and Y co-ordinates of edge. Position of Centroid i.ecentre of edge component (CX, CY) is computed as,

$$Cx = \frac{1}{N}\sum_{i=1}^{n} Xi\ \ and\ \ Cy = \frac{1}{N}\sum_{i=1}^{n} Yi \qquad (3.1)$$

To calculate distance between Centroid and edge component, distance is computed as,

$$Distance = SQRT(abs(CX - X^1)^2 +\ abs(CY - Y^1)^2) \qquad (3.2)$$

$$(X^1, Y^1 = (Cx + A, Cy + B) and\ A, B = \{-1..1\} \qquad (3.3)$$

$$PSC = \frac{Number\ of\ Straight\ components}{Total\ number\ of\ \ components} \qquad (3.4)$$

$$PCC = \frac{Number\ of\ Cursive\ components}{Total\ number\ of\ \ components} \qquad (3.5)$$

The above steps are illustrated in Fig. 3.3-Fig. 3.8 respectively for Indus, Kannada, Tamil, Telugu and Malayalam document images. It is observed from Fig. 3.3(d) that the number of components which satisfy the straightness property are less because the centroid does not fall on the same document for most of the components while from

Fig. 3.4(d)-Fig. 3.8(d), most of the components satisfy straightness property because the centroid of the components fall on the same components for most of the components. Therefore, it is concluded that the percentage of straightness is lower than the percentage of cursiveness for Indus document and it is vice versa for other documents.

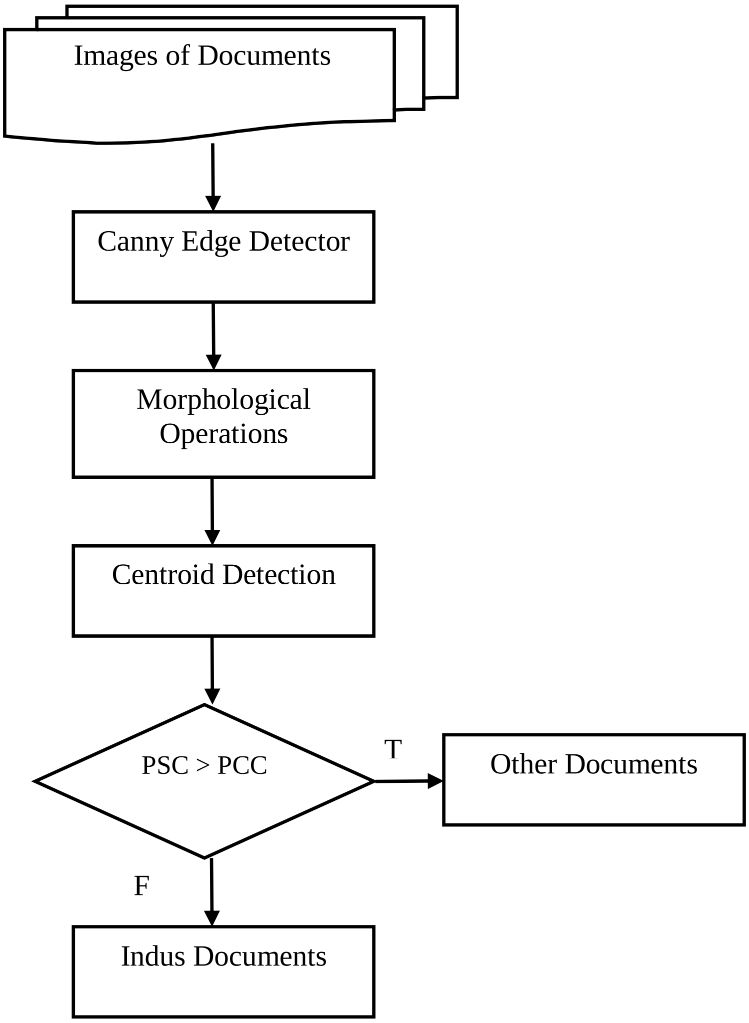
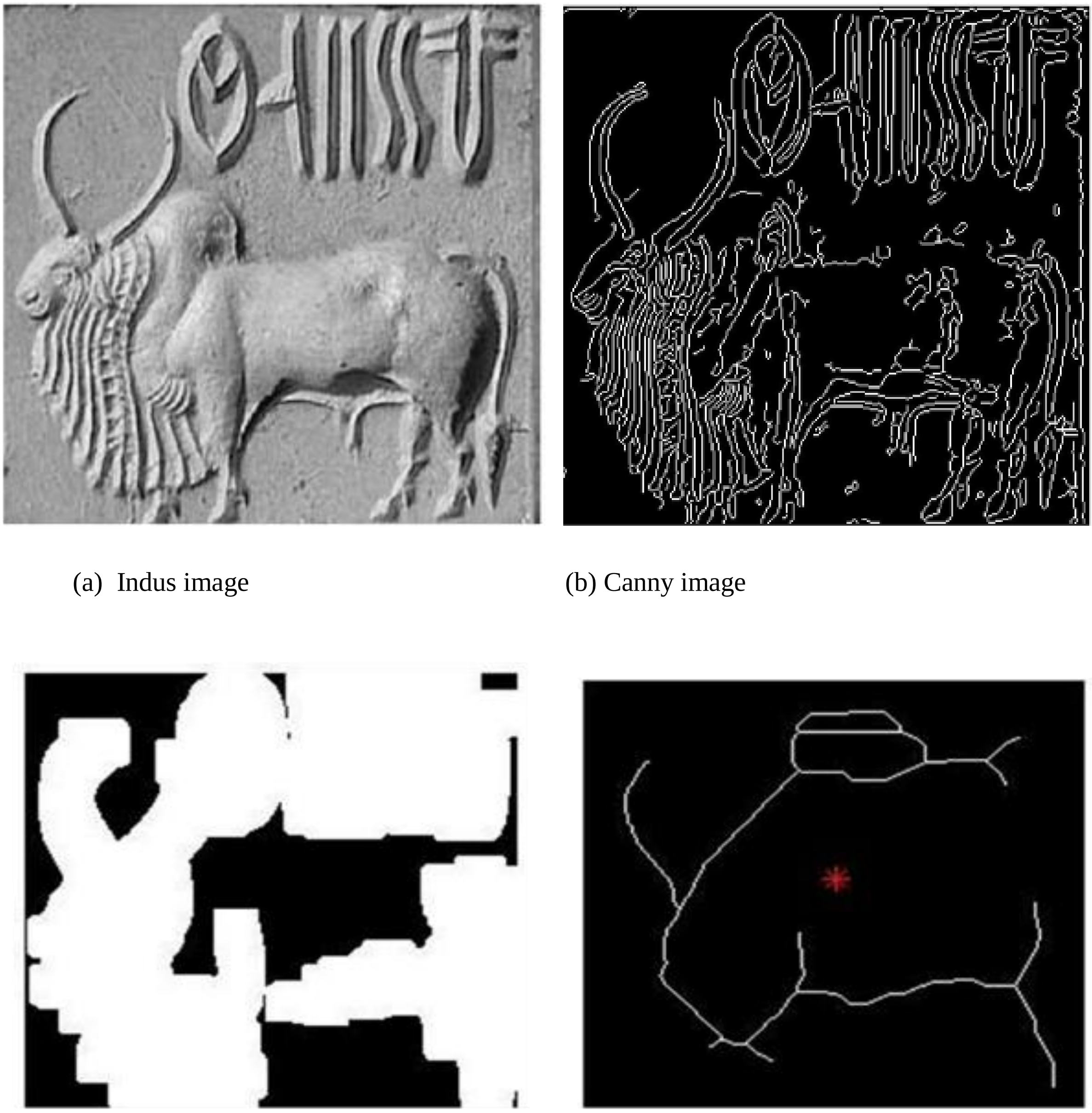


Fig. 3.2. Flow diagram of the Centroid based Method



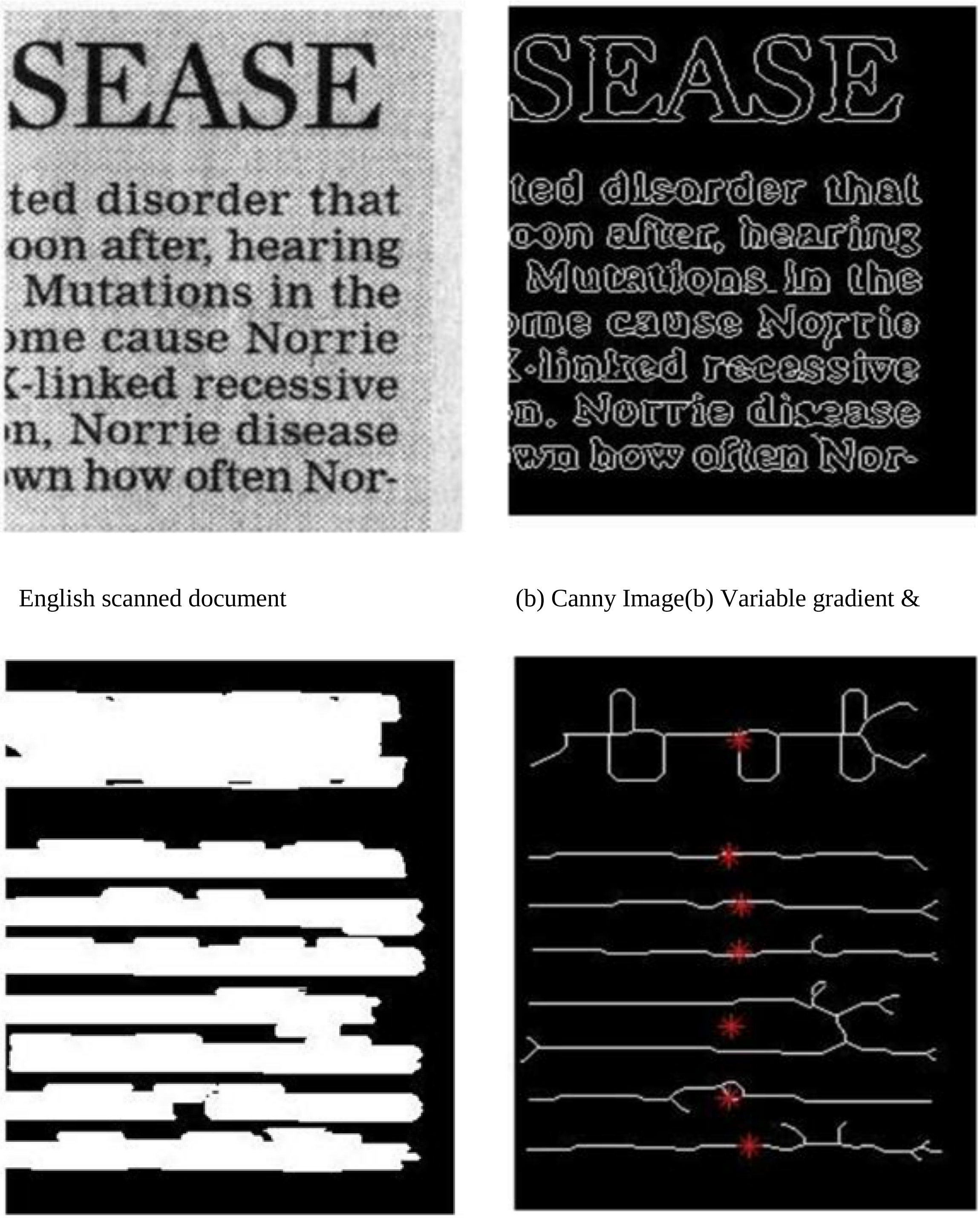
(a) Indus image

(b) Canny image

(c) Dilated image

(d) Thinned image (Centroid marked)

Fig. 3.3.Morphological operations on Indus document image



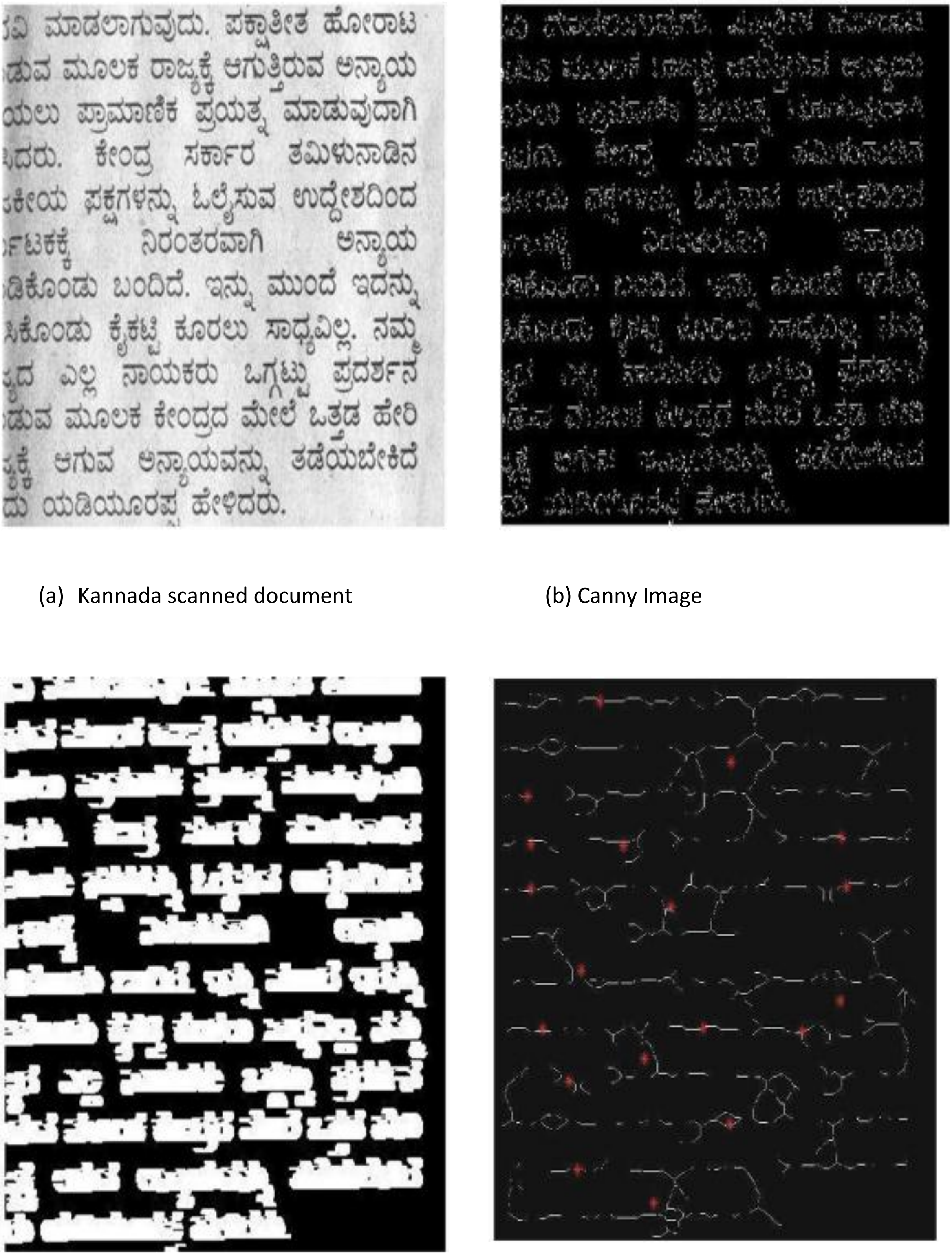
English scanned document

(b) Canny Image(b) Variable gradient &

(c) Dilated image

(d) Thinned image (Centroid marked)

Fig. 3.4.Morphological operations on English document image



ವಿ ಮಾಡಲಾಗುವುದು. ಪಕ್ಷಾತೀತ ಹೋರಾಟ
ಡುವ ಮೂಲಕ ರಾಜ್ಯಕ್ಕೆ ಆಗುತ್ತಿರುವ ಅನ್ಯಾಯ
ಯಲು ಪ್ರಾಮಾಣಿಕ ಪ್ರಯತ್ನ ಮಾಡುವುದಾಗಿ
ದರು. ಕೇಂದ್ರ ಸರ್ಕಾರ ತಮಿಳುನಾಡಿನ
ಕೀಯ ಪಕ್ಷಗಳನ್ನು ಓಲೈಸುವ ಉದ್ದೇಶದಿಂದ
ಟಕಕ್ಕೆ ನಿರಂತರವಾಗಿ ಅನ್ಯಾಯ
ಡಿಕೊಂಡು ಬಂದಿದೆ. ಇನ್ನು ಮುಂದೆ ಇದನ್ನು
ಸಿಕೊಂಡು ಕೈಕಟ್ಟಿ ಕೂರಲು ಸಾಧ್ಯವಿಲ್ಲ. ನಮ್ಮ
ದ ಎಲ್ಲ ನಾಯಕರು ಒಗ್ಗಟ್ಟು ಪ್ರದರ್ಶನ
ಡುವ ಮೂಲಕ ಕೇಂದ್ರದ ಮೇಲೆ ಒತ್ತಡ ಹೇರಿ
ಕ್ಕೆ ಆಗುವ ಅನ್ಯಾಯವನ್ನು ತಡೆಯಬೇಕಿದೆ
ದು ಯಡಿಯೂರಪ್ಪ ಹೇಳಿದರು.

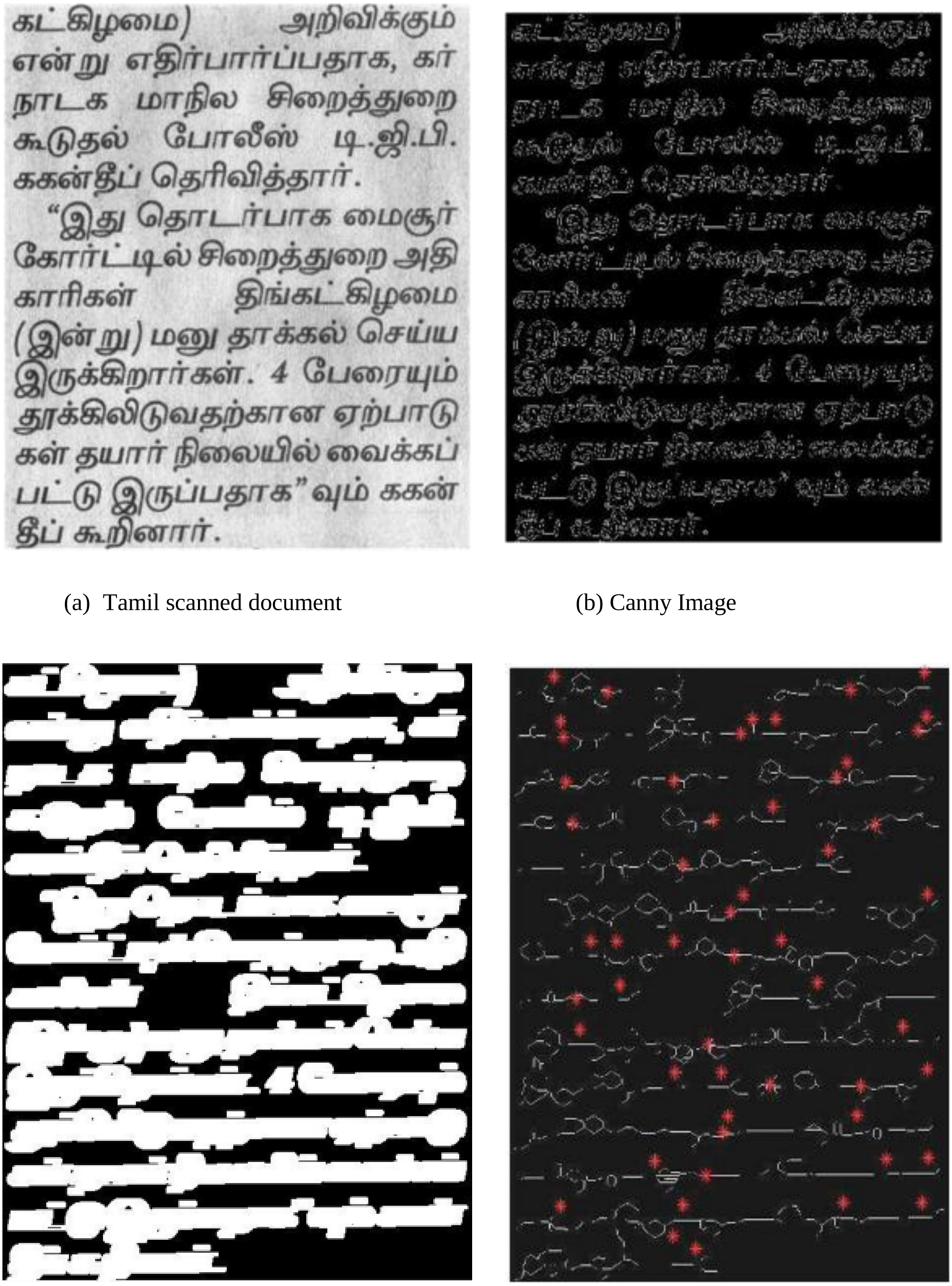
(a) Kannada scanned document

(b) Canny Image

(c) Dilated image

(d) Thinned image (Centroid marked)

Fig. 3.5.Morphological operations on Kannada document image



கட்கிழமை) அறிவிக்கும்
என்று எதிர்பார்ப்பதாக, கர்
நாடக மாநில சிறைத்துறை
கூடுதல் போலீஸ் டி.ஜி.பி.
ககன்தீப் தெரிவித்தார்.
"இது தொடர்பாக மைசூர்
கோர்ட்டில் சிறைத்துறை அதி
காரிகள் திங்கட்கிழமை
(இன்று) மனு தாக்கல் செய்ய
இருக்கிறார்கள். 4 பேரையும்
தூக்கிலிடுவதற்கான ஏற்பாடு
கள் தயார் நிலையில் வைக்கப்
பட்டு இருப்பதாக" வும் ககன்
தீப் கூறினார்.

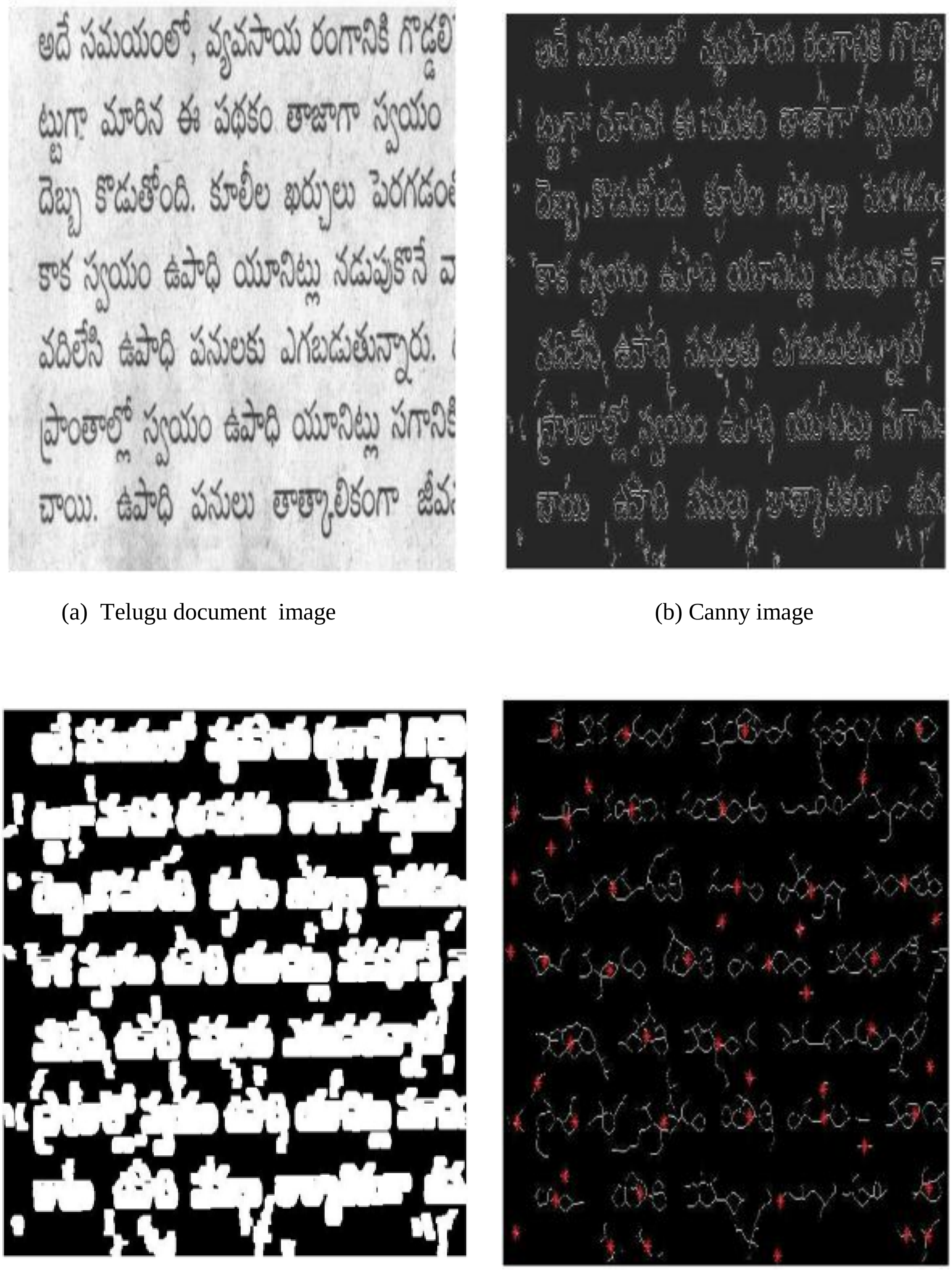
(a) Tamil scanned document

(b) Canny Image

(c) Dilated image

(d) Thinned image (Centroid marked)

Fig. 3.6.Morphological operations on Tamil document image


అదే సమయంలో, వ్యవసాయ రంగానికి గొడ్డలి
ట్టుగా మారిన ఈ పథకం తాజాగా స్వయం
దెబ్బ కొడుతోంది. కూలీల ఖర్చులు పెరగడం
కాక స్వయం ఉపాధి యూనిట్లు నడుపుకొనే వ
వదిలేసి ఉపాధి పనులకు ఎగబడుతున్నారు.
ప్రాంతాల్లో స్వయం ఉపాధి యూనిట్లు సగానికి
చాయి. ఉపాధి పనులు తాత్కాలికంగా జీవ

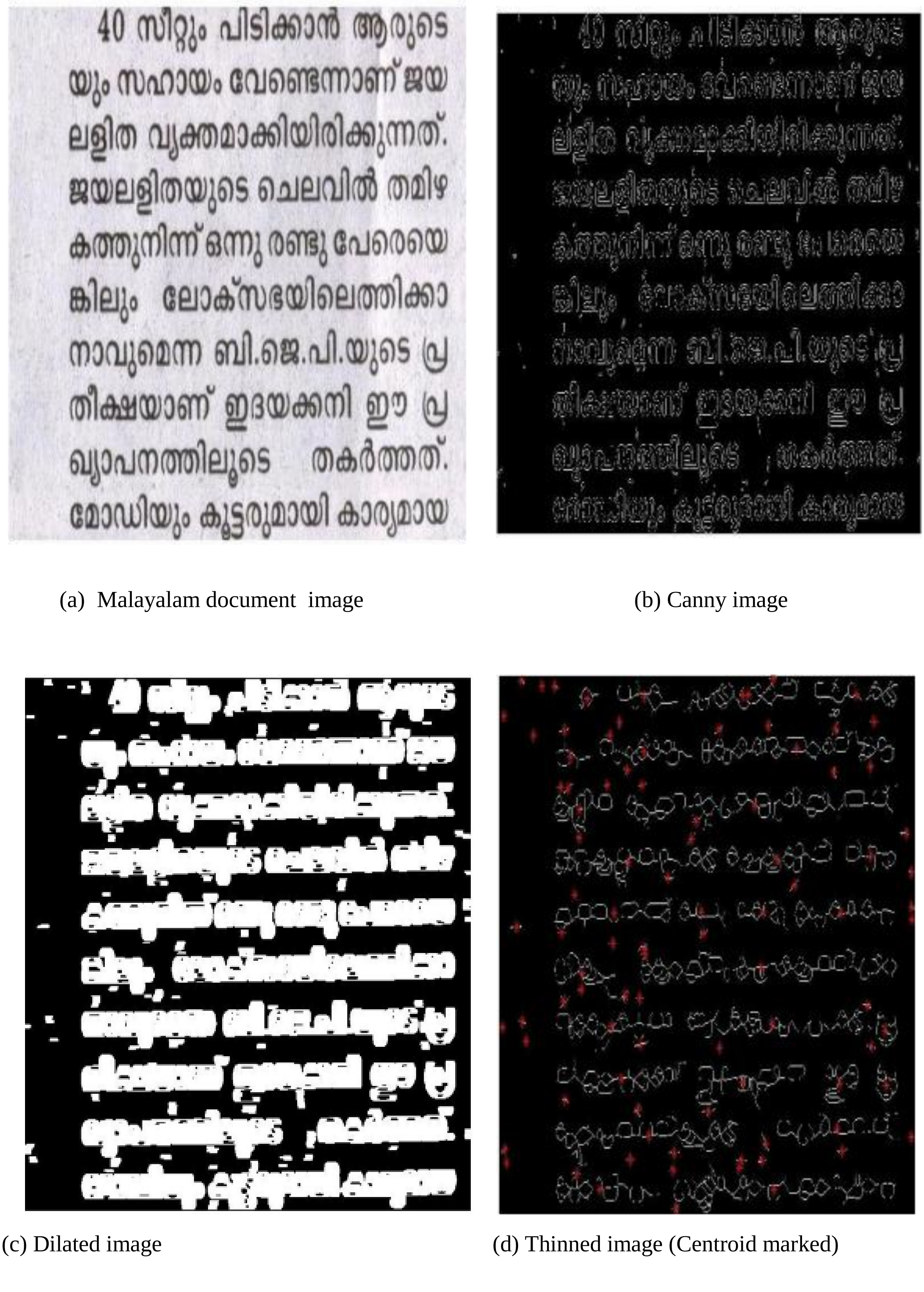
(a) Telugu document image

(b) Canny image

(c) Dilated image

(d) Thinned image (Centroid marked)

Fig. 3.7.Morphological operations on Telugu document image



40 സീറ്റും പിടിക്കാൻ ആരുടെ
യും സഹായം വേണ്ടെന്നാണ് ജയ
ലളിത വ്യക്തമാക്കിയിരിക്കുന്നത്.
ജയലളിതയുടെ ചെലവിൽ തമിഴ
കത്തുനിന്ന് ഒന്നു രണ്ടു പേരെയെ
ങ്കിലും ലോക്സഭയിലെത്തിക്കാ
നാവുമെന്ന ബി.ജെ.പി.യുടെ പ്ര
തീക്ഷയാണ് ഇദയക്കനി ഈ പ്ര
ഖ്യാപനത്തിലൂടെ തകർത്തത്.
മോഡിയും കൂട്ടരുമായി കാര്യമായ

(a) Malayalam document image

(b) Canny image

(c) Dilated image

(d) Thinned image (Centroid marked)

Fig. 3.8.Morphological operations on Malayalam document image

3.2.1.2 Region Growing Based Straightness and Cursiveness

The algorithm presented in previous section involves morphological information to smooth the components and merge disconnected component as one component to study the structure of the components. Since morphological operation depends on mask size, it affects shape of the components especially for the components of different font and font size. As a result, the above algorithms classifies English document as south Indian documents. To overcome this problem, we propose region growing based procedure for merging disconnected components into single component which preserve the structure of the components. The proposed method uses the results of region growing instead of the results of morphological operation used in the previous algorithm for script identification. In other words, for the algorithm presented in previous section uses region growing instead of morphological operation for calculating percentage of straight and cursive components to identify the scripts. It is noted that usually components in south Indian documents are more cursive compared to the components in English document because of modifiers present in the south Indian documents. As a result, one can expect more cursive components for south Indian documents and more straight components for English documents. This is the basis for the proposed method to identify the English script from south Indian scripts. The steps of the algorithm are shown in Fig. 3.9.

The proposed method fixes bounding boxes for the components in the edge image of the input image. Then the method allows to grow boundary of the component pixel by pixel in horizontal direction until it reaches nearest neighbor bounding box of the component. In this way, the proposed method merges neighbor components to get one component for the text line. Then the method performs thinning as in the previous algorithm to obtain single pixel width components. Sample results for the above steps for the English and south Indian documents are shown from Fig. 3.10.

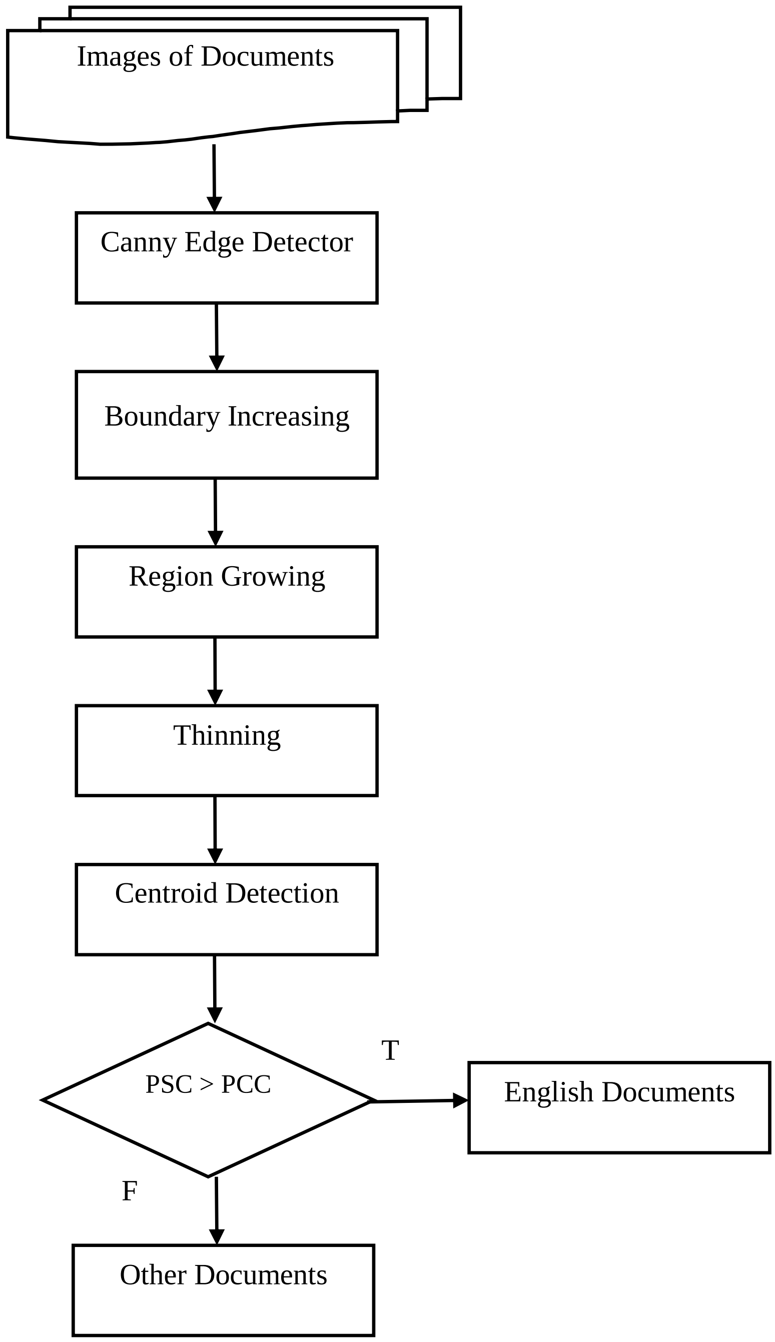
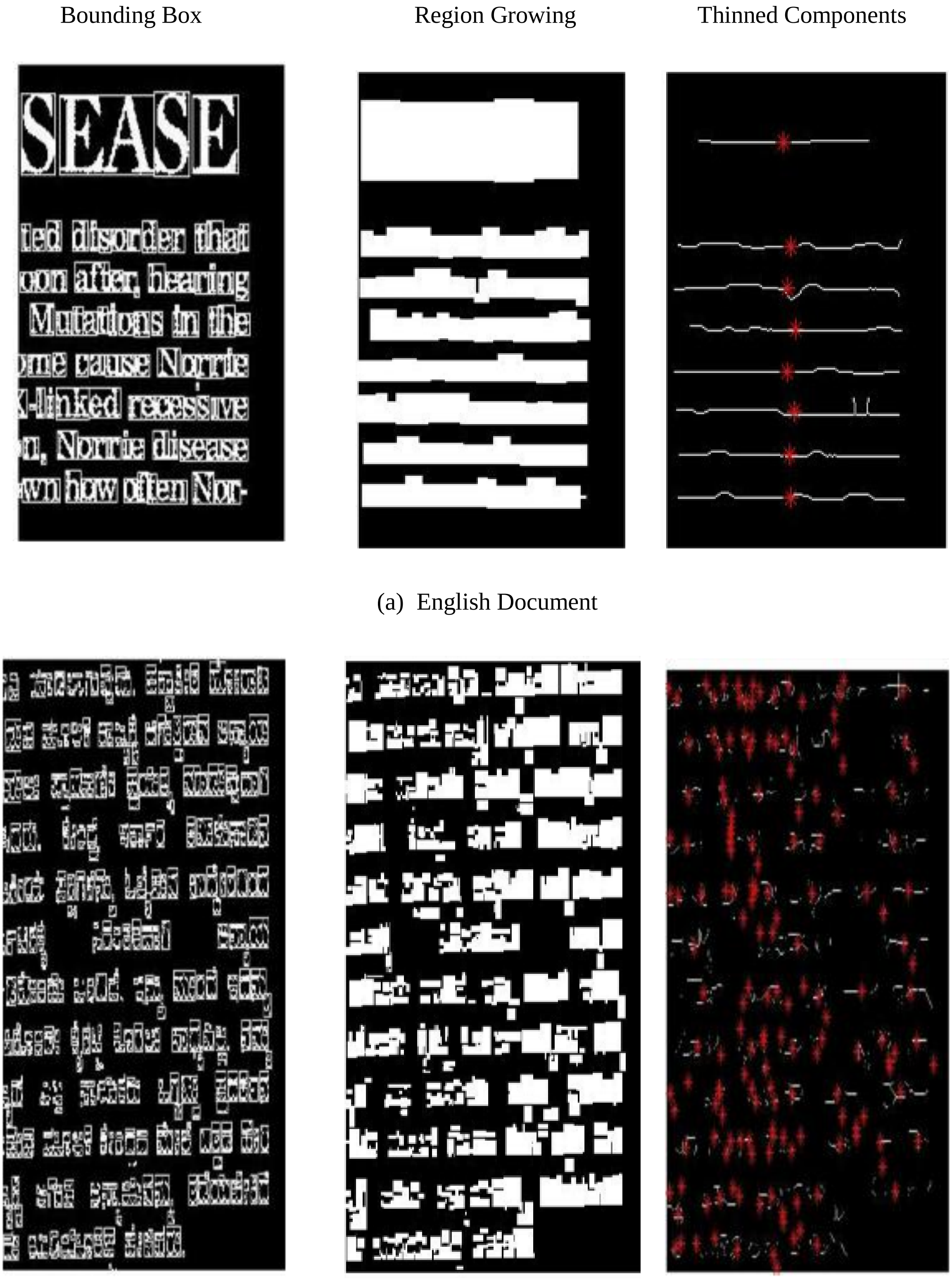


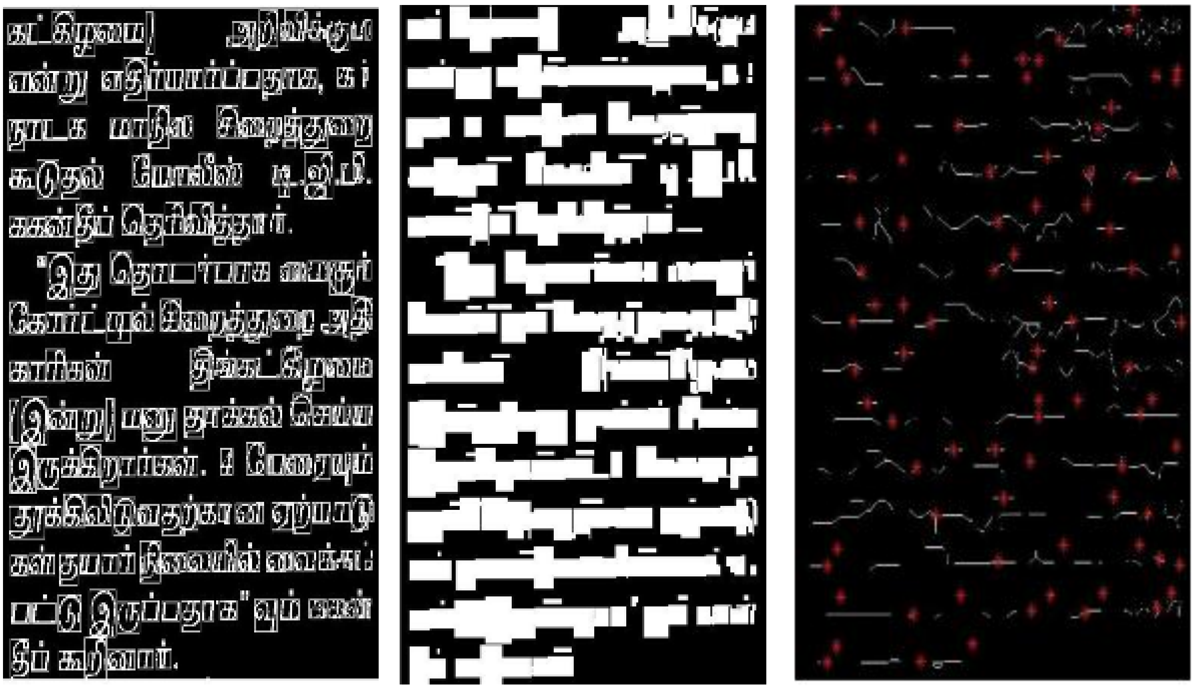
Fig. 3.9. Flow diagram of the Region Growing based Method



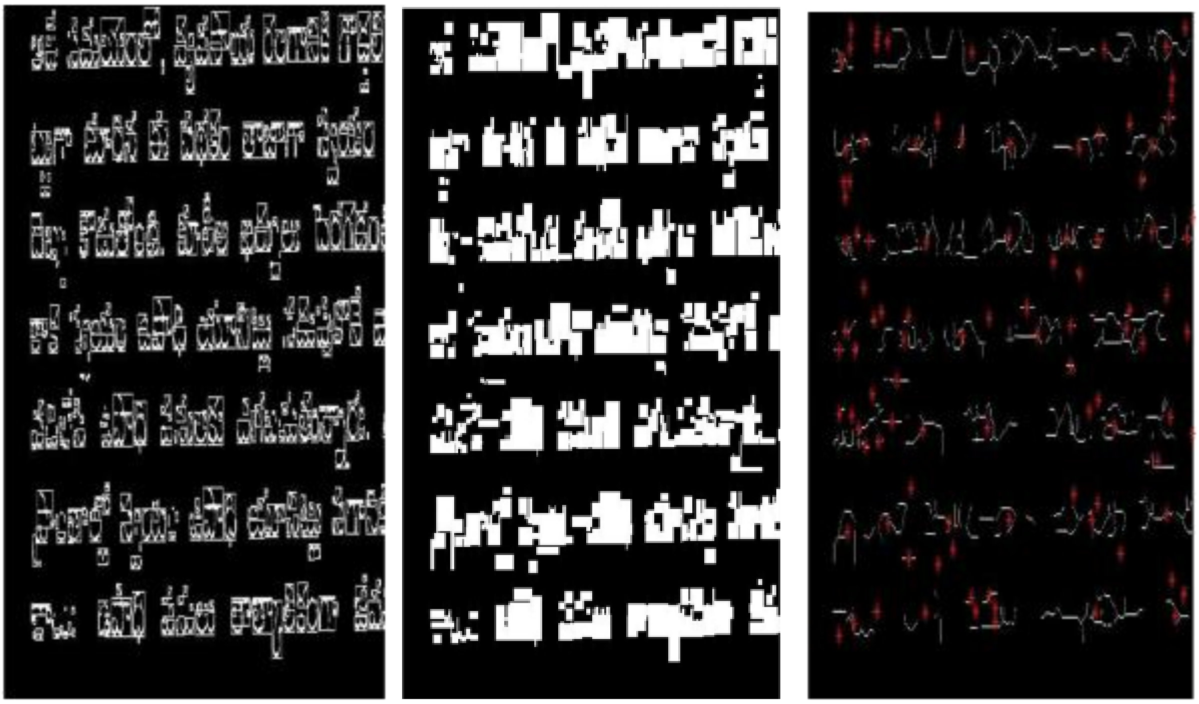
Bounding Box Region Growing Thinned Components

(a) English Document

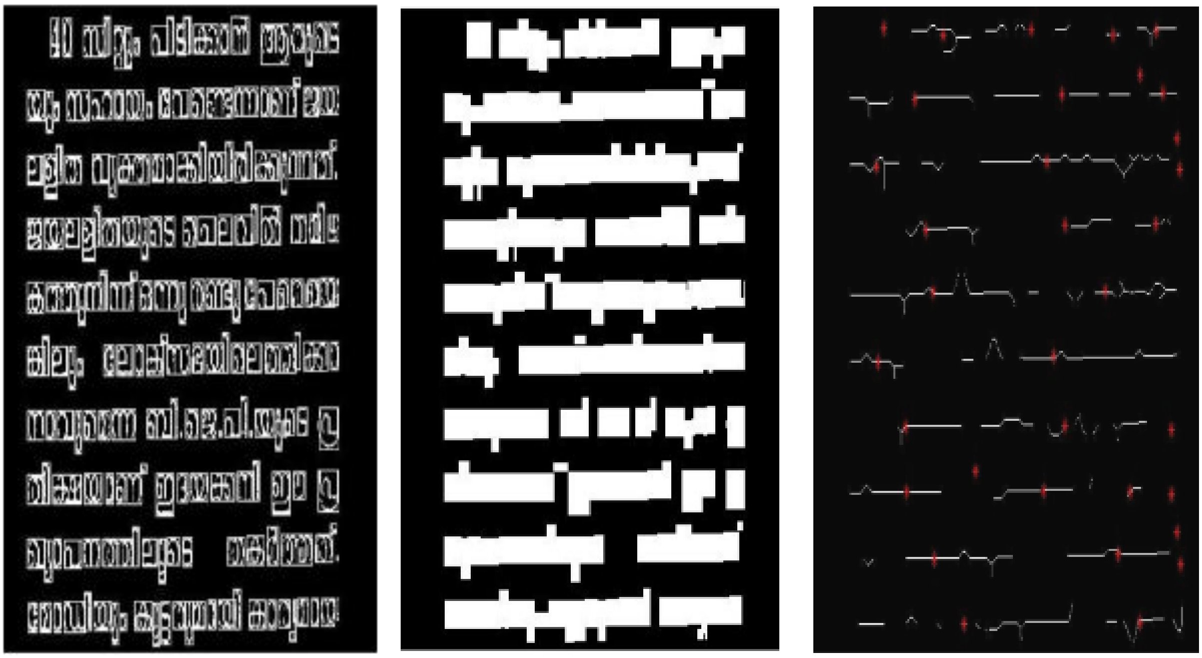
(b) Kannada Document



Tamil Document



(c) Telugu Document



(d) Malayalam Document

Fig. 3.10.Straight/ Cursive components detection using Region growing method

3.2.2 An Approach based on Skew and Orientation of the Components

The algorithms presented in Section 3.2.1 and Section 3.2.2 is not robust to degradations. Therefore, we propose a method based on orientation of the components rather than checking the condition that centroid falls on component or not for script identification in this section. This section presents two methods, one is finding slope for the region which uses morphological operation and another uses nearest neighbor criteria for merging the components. The former one is good to classify Indus from other document which includes four south Indian and English and latter one classifies English from south Indian documents.

3.2.2.1 Morphological Region Slope based Straightness and Cursiveness

The method follow the same steps presented in Section 3.2.1 except checking the condition whether centroid falls on the same component to classify it as straight or cursive component. In this algorithm, we calculate angle for the thinned components. The average of the angles of all the components is estimated. The proposed method compares angle of each component with average angle of the same component. If angle of the component is more than certain threshold then component is considered as straight components else cursive components. Once it classifies straight and cursive components using angle information, the proposed method calculates percentage as algorithm in Section 3.2.1.1 for both straight and cursive components. If the percentage of the cursive components of the document is larger than the percentage of straight components, document classified as Indus document else other documents as algorithm in Section 3.2.1.1 The steps of the algorithm are shown in Fig. 3.11. Sample results for the steps are shown in Fig. 3.12 for Indus document and Fig. 3.13 for South Indian, and English documents.

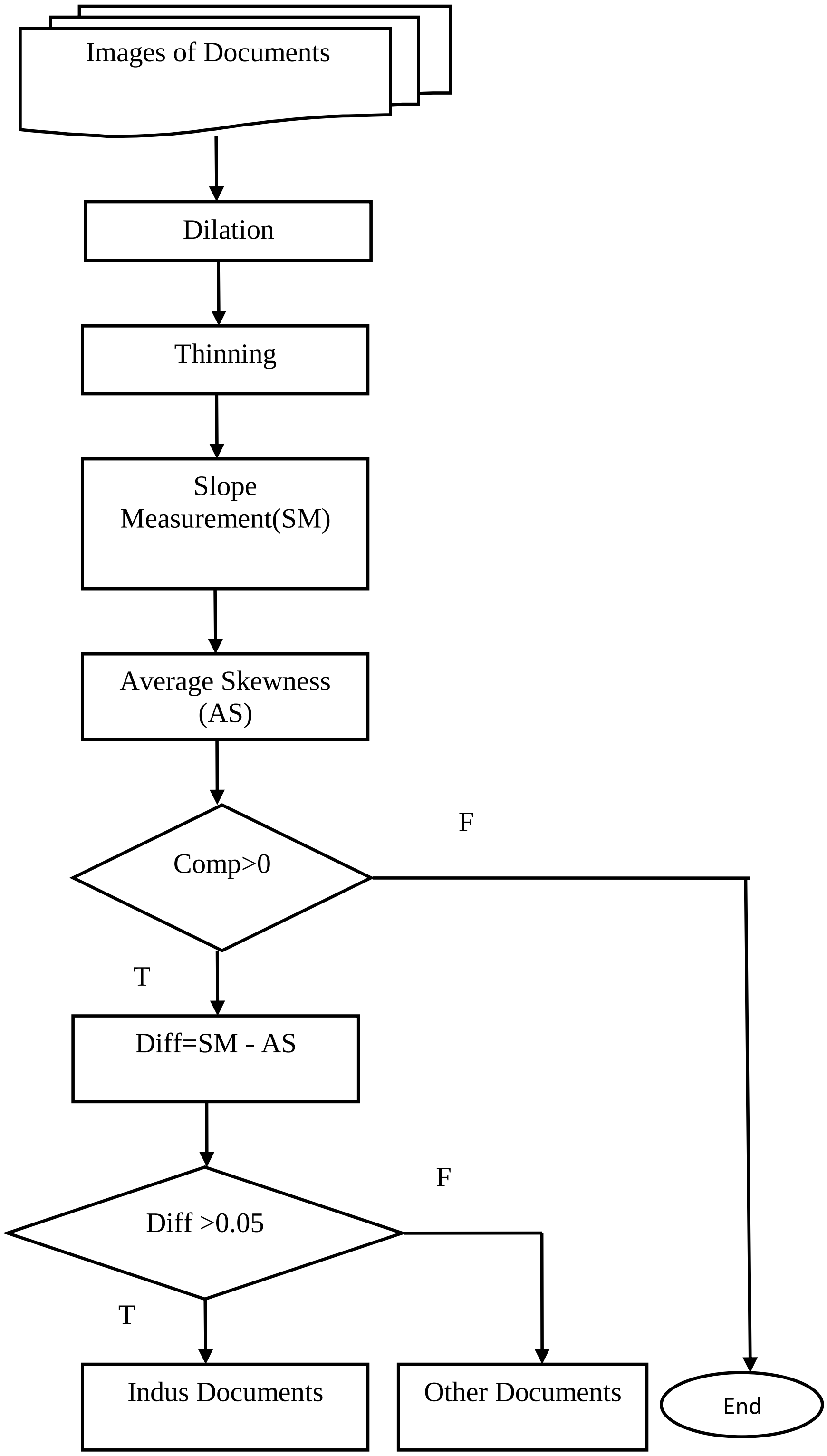


Fig. 3.11. Flow diagram of the Region Slope based Method

Mathematically, skewness of the component is measured as follows. Let (x1,y1) and (x2,y2) are the extreme coordinates of thinned component, angle of each component is defined as the difference of y coordinates, say (y2 - y1) of the component divided by the extreme x2 of the thinned component. Mathematically, the angle of each connected component is defined as follows. Let X={x1, x2} and Y= {y1, y2} be the startand end coordinates of an edge, respectively.

The Skewness of the component, $Ang(cc)$ computed as,

$$Ang(cc) = \frac{y2 - y1}{x2 - x1} \quad (3.6)$$

$$Ang(cc) = \frac{y2 - y1}{x2} \quad (3.7)$$

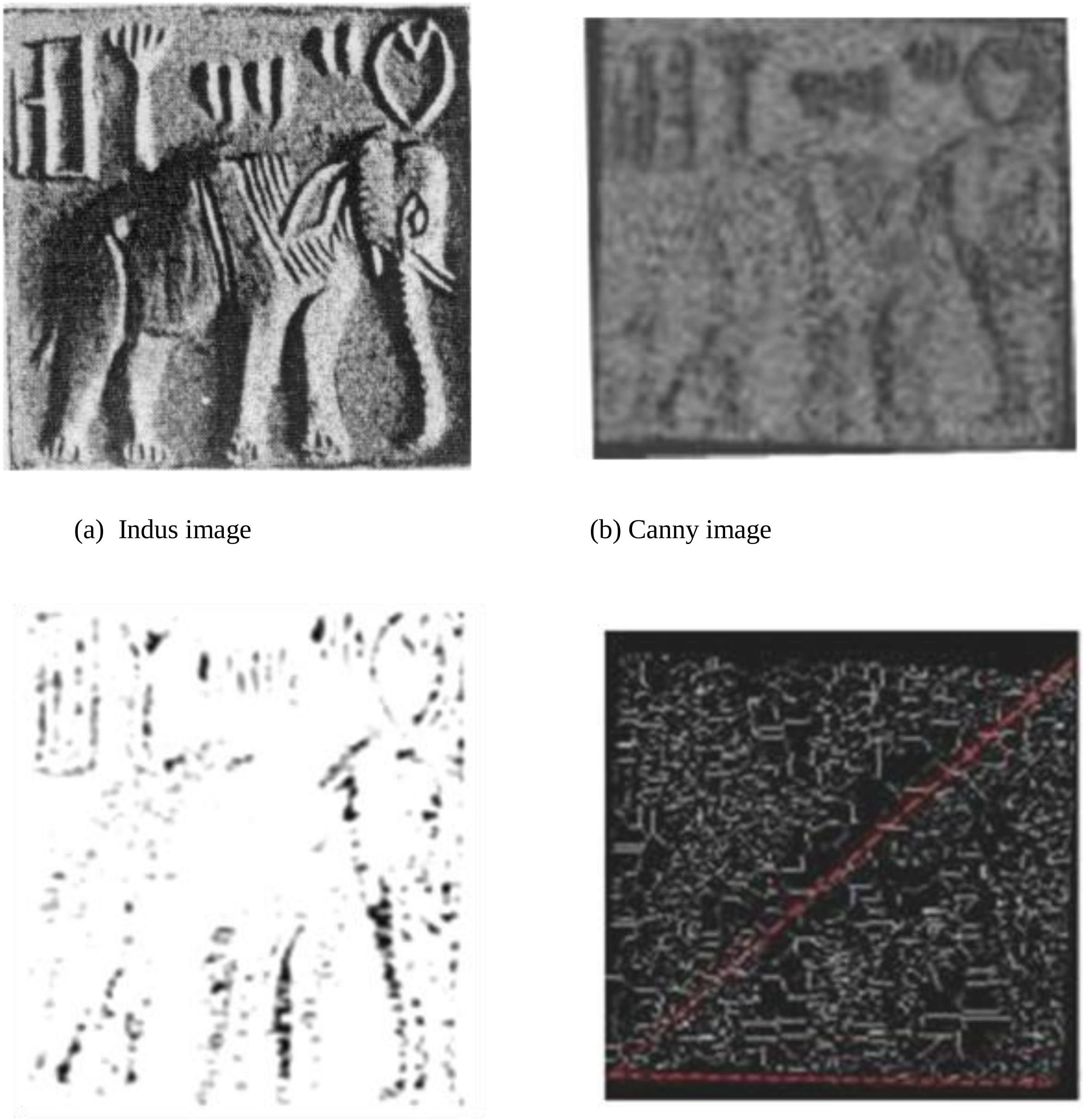
$$Avg = 1/n \sum_{i=1}^{n} Ang(cci) \quad (3.8)$$

For identification of near or far components, the method finds the difference of average angle and the skewness of each component as in equation (3.9). Each edge component is considered as far if the difference is greater than certain range (0.05 radii). Edge component is considered as near if the difference is lesser or equal to 0.05 radii. Image is considered as Indus if percentage of far components(PFC) is more than percentage of near components(PNC). Image is considered as others (English or other documents) if percentage of near components $(PFC)$are more than percentage of far components(PNC).

$$Diff(cci) = Avg - Ang(cci) \quad (3.9)$$

$$PFC = \frac{Number\ of\ far\ components}{Total\ number\ of\ \ components} \quad (3.10)$$

$$PNC = \frac{Number\ of\ near\ components}{Total\ number\ of\ \ components} \tag{3.11}$$



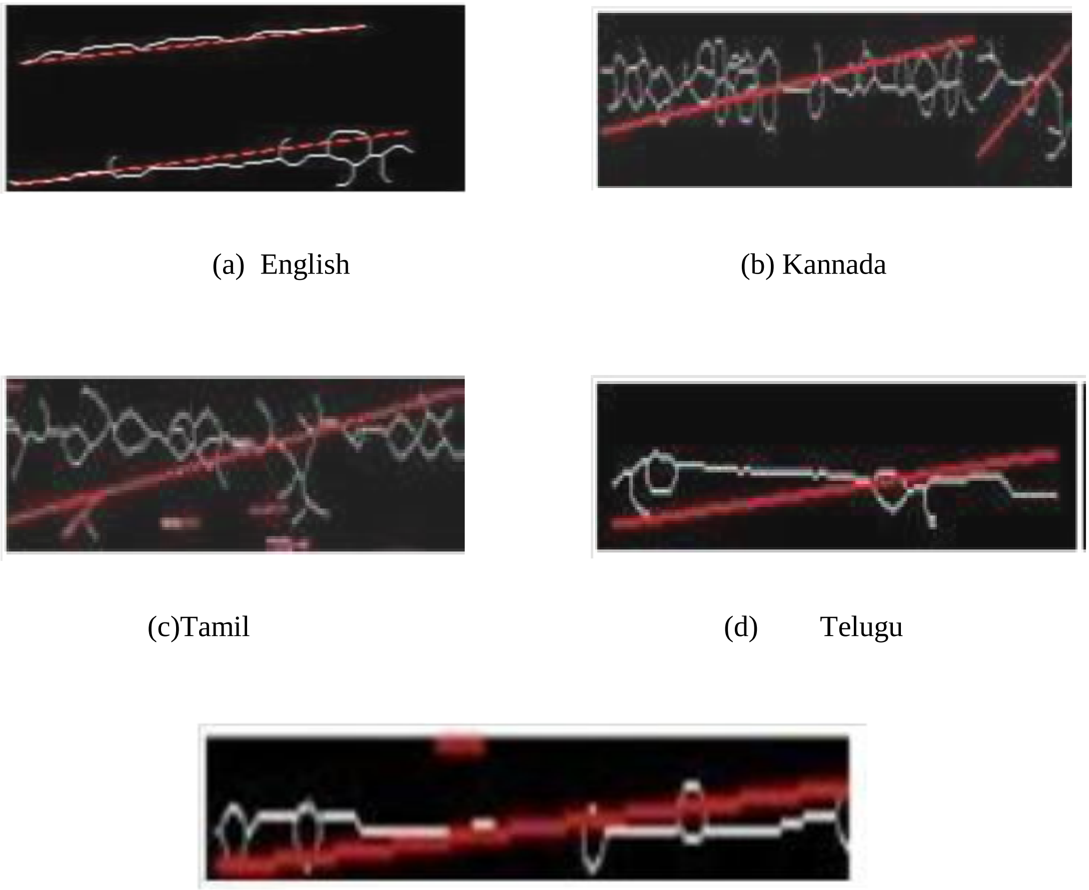
(a) Indus image

(b) Canny image

(c) Dilated image

(d) Thinned image

Fig. 3.12.Morphological operations on Indus document image



(a) English

(b) Kannada

(c)Tamil

(d) Telugu

(e) Malayalam

Fig. 3.13.Skewness measurement in the components of English and South Indian scripts

3.2.2.2 Nearest Neighbor Clustering based Straightness and Cursiveness

As noted from algorithm presented in Section 3.2.2.1, morphological operation based grouping component may not preserve structure of the components, we propose nearest neighbor clustering as mentioned in section 3.2.2.2 for grouping the components. The above algorithm is good for classifying Indus from south Indian and English documents. This algorithm is developed to classify English from south Indian documents. The algorithms follows the same steps as presented in previous section except grouping components based on nearest neighbor clustering to calculate angle. In other words, after grouping using nearest neighbor clustering, the method uses the same steps to identify modifiers which present in south Indian documents. If the modifier present in the component then the component is considered as cursive component else straight component. After this step, the method calculates percentage of straight and cursive components for the document identify the script. This method preserves structure in better way compared to the previous algorithm because it works based on the fact that the distance between character components is less than the distance between text lines. As a results, nearest neighbor clustering helps in finding modifiers which present in south Indian documents. Sample results for each step can be seen in Fig. 3.14 where one can see the algorithm identifies modifiers based on distance for the Kannada text line. Therefore, south Indian documents gets high percentage of cursive components compared to straight components. Since English document does not provide modifiers, the percentage of straight components are higher than cursive components. Distance between the components is estimated as, Assume there are three components $C1, C2, C3$. Where $C1, C2$ the character component is $C3$is the modifier component.

$$C1 = \{(x11, y11); (x12, y12); (x13, y13) \ldots (x1n, y1n)\} \quad (3.12)$$

$$C2 = \{(x21, y21); (x22, y22); (x23, y23) \ldots (x2n, y2n)\} \quad (3.13)$$

$$C3 = \{(x31, y31); (x32, y32); (x33, y33) \ldots (x3n, y3n)\} \quad (3.14)$$

Distance between character components is computed as,

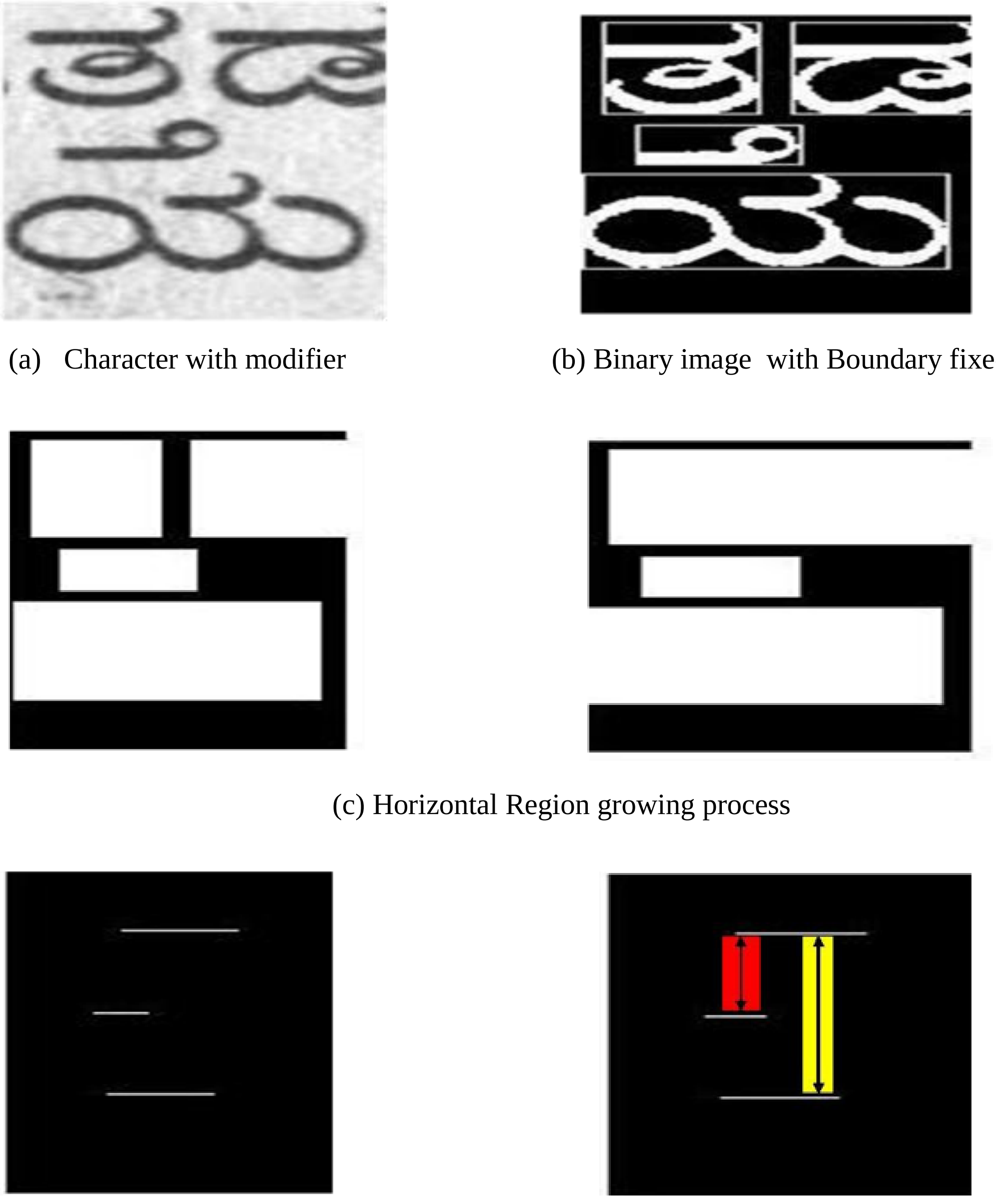
$$Dist1 = (\forall i = 1..n), \min(sqrt(x1i - x2i)^2 + (y1i - y2i)^2) \quad (3.15)$$

Distance between character component and modifier component is computed as,

$$Dist2 = (\forall i = 1..n), \min(sqrt(x3i - x1i)^2 + (y3i - y1i)^2)\ ) \quad (3.16)$$

$$Height = Rowmax - Rowmin\ and\ Width = Colmax - Colmin \quad (3.17)$$

If $Dist1 < Dist2$ then the method considers the given component as modifiers. If the amounts of modifiers are more then it means cursiveness of the script is more. More the cursiveness then the document is considered as South Indian scripts. If $Dist2 = 0$ then the component does not have modifiers. More the amount of components without modifiers then the document is considered as English scripts. Hence the Nearest Neighbor Clustering method groups the horizontal components and checks the nearest component vertically to check the presence of modifiers.



(a) Character with modifier

(b) Binary image with Boundary fixed)

(c) Horizontal Region growing process

(d) Thinning

(e) Distance estimation

Fig. 3.14. NNA for identifying Modifiers / Diacritics

3.2.3 An Approach based on Spatial Relationship of Dominant Points of the Components

The algorithms presented in previous sections are limited to Indus, English and South Indian documents. The algorithm fails to classify south Indian documents as Kannada, Tamil, Telugu and Malayalam. This is because the features are good enough to classify. Therefore, in this section, we present a new method for classifying Indus, English, Kannada, Tamil,Telugu, Hindi and Gujarati. It is true that each script has its own alphabets and shapes. The method exploits the shapes of the components to classify the different scripts in this section. To extract unique shape of the components of different scripts, we extract dominant points of edge component, namely, junction, intersection and end points and then the method finds spatial relationship between the dominant points to identify the scripts. Flow diagram of the entire proposed method is shown in Fig.3.15.

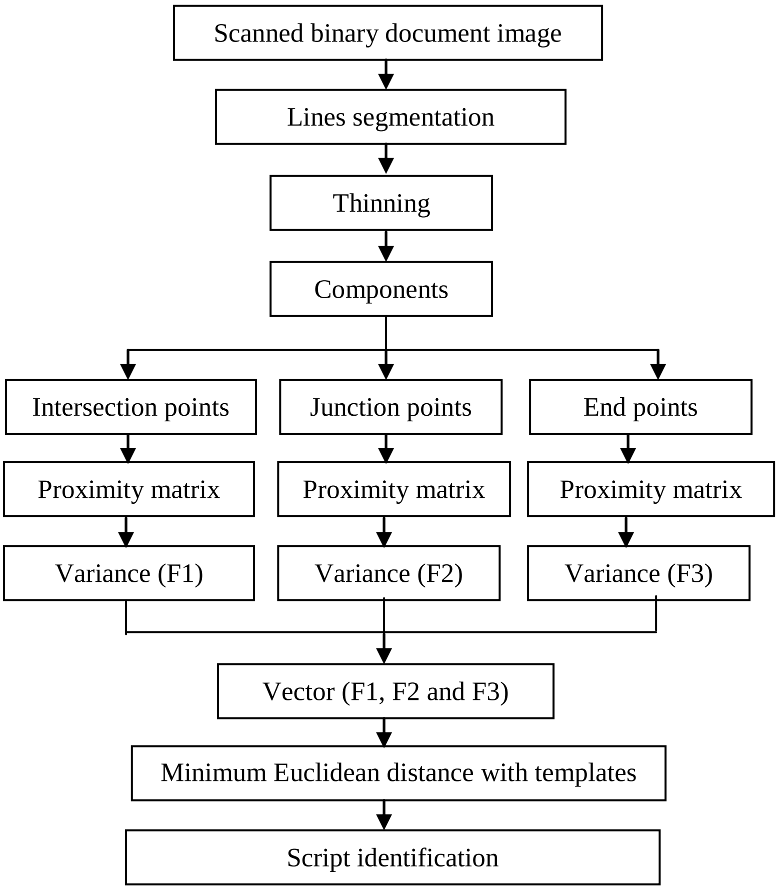


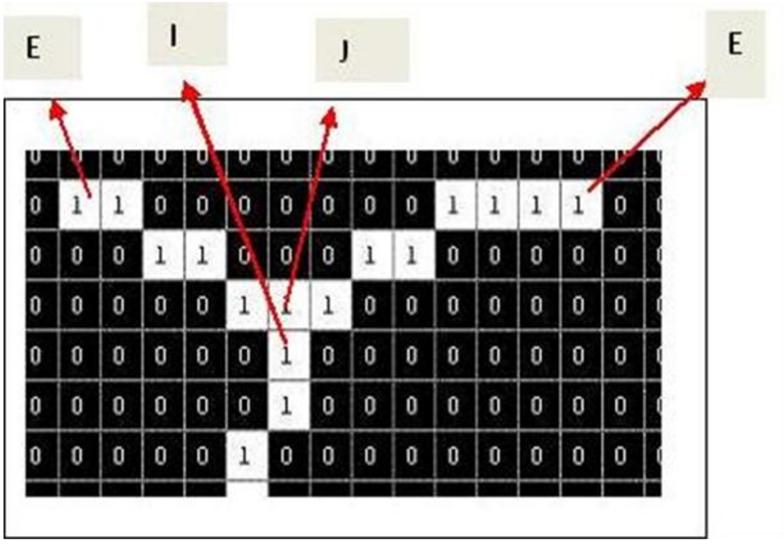
Fig. 3.15. Logical flow of the proposed method

3.2.3.1 Dominant Pixel Detection

Proposed method converts scanned gray image to binary. Therefore, to extract the dominant points the method reduces the width of edge pixels to single pixel widthby applying thinning algorithm. For each thinned edge component, the method determines intersection, Junction and End point as in Fig.3.16. We define Intersection point as, the point at which its removal give rise to four independent edges and such pixels have 4 neighbors' pixels. Junction point is the point at which its removal gives rise to three independent edges and it has 3 neighbor pixels. End point is the point at which it has one neighbor pixel. To determine dominant points such as Intersection, Junction and End points, the method tracks each of the component regions using bounding box and finds the height and width of each component using the positions of boundary pixels.Four different positions evaluated from bounding box are, Rowmin, Rowmax, Colmin, and Colmax. These pixel positions are found as follows.

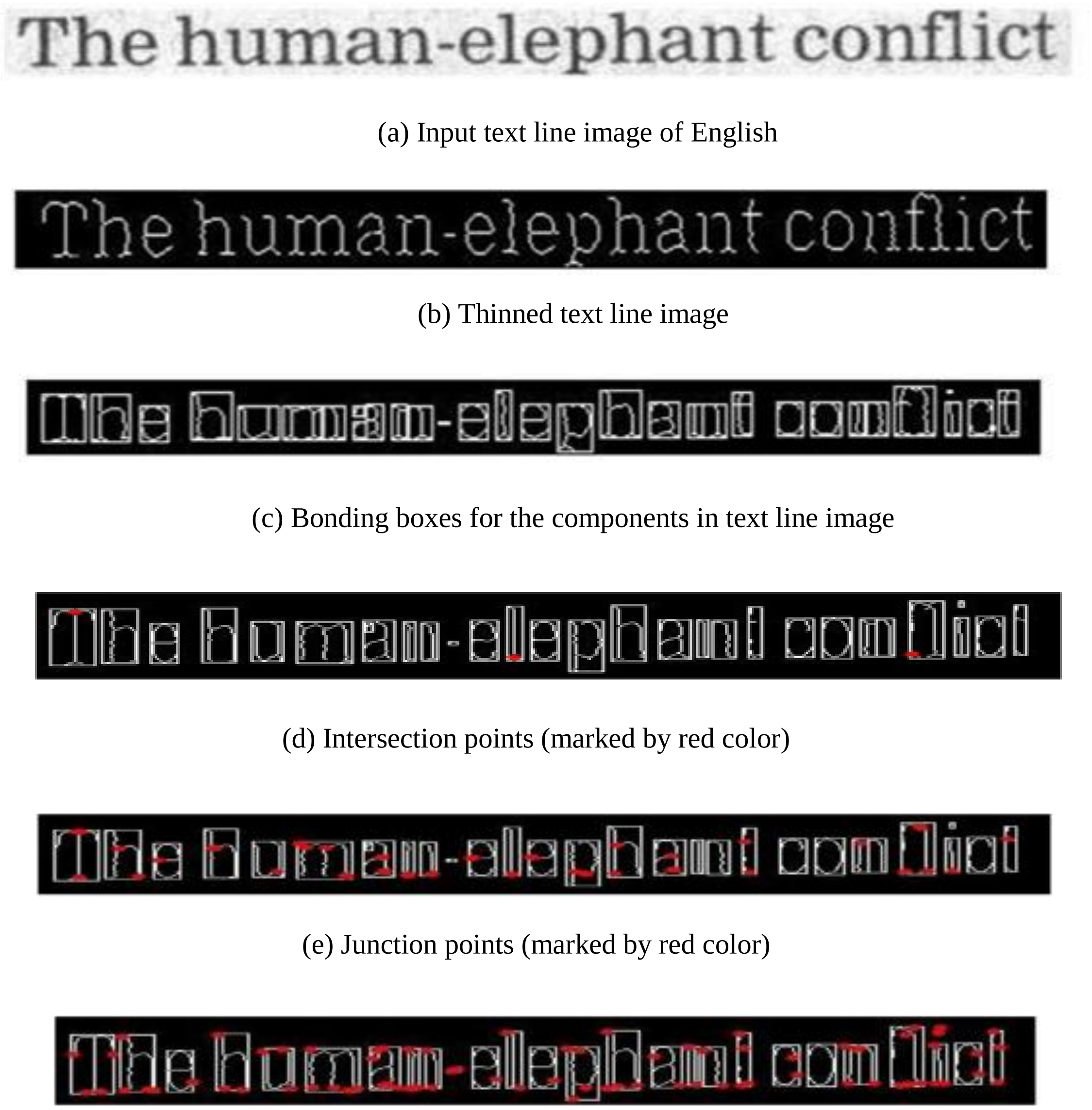
- Rowmin: This is the position of firstwhite pixel found in first row of connected component.
- Rowmax: This is the white pixel found in last row of connected component.
- Colmin: This is pixel position correspond to first column and first row between Rowmin and Rowmax.
- Colmax: This is pixel position correspond to last column and first row between Rowmin and Rowmax.

To detect various dominant points such as, Intersection, Junction and End points the method, checks the number of neighbors for each pixel in the component within the boundary determined. To define boundary extremes best fitting rectangle is detected for each connected component as shown in Fig 3.17(c). Height and width of bounding box is computed as in equation 3.17. Dominant pixels detection process for entire text line of English script is shown in Fig.3.17 (d-f). Where it is seen very less number of intersection (Fig.3.17 (d)), junction (Fig.3.17 (e)) and End points (Fig.3.17 (f)) compare to other scripts.In the same way dominant pixels are detected for different script text lines. Dominant pixel detection based on count of neighbors for various scripts is shown in Fig.3.18. Dominant pixels detected for sample of Indus document image is shown in Fig.3.18 (a). Dominant pixels detected for word samples of different languages shown in Fig.3.18 (b-g).In this way dominant pixels are extracted from each of the scripts.



0	1	1	0	0	0	0	0	0	0	1	1	1	1	0
0	0	0	1	1	0	0	0	1	1	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Fig. 3.16. Dominant Pixels selection for each component: I- Intersection points, J- Junction points, E-End points.



(a) Input text line image of English

(b) Thinned text line image

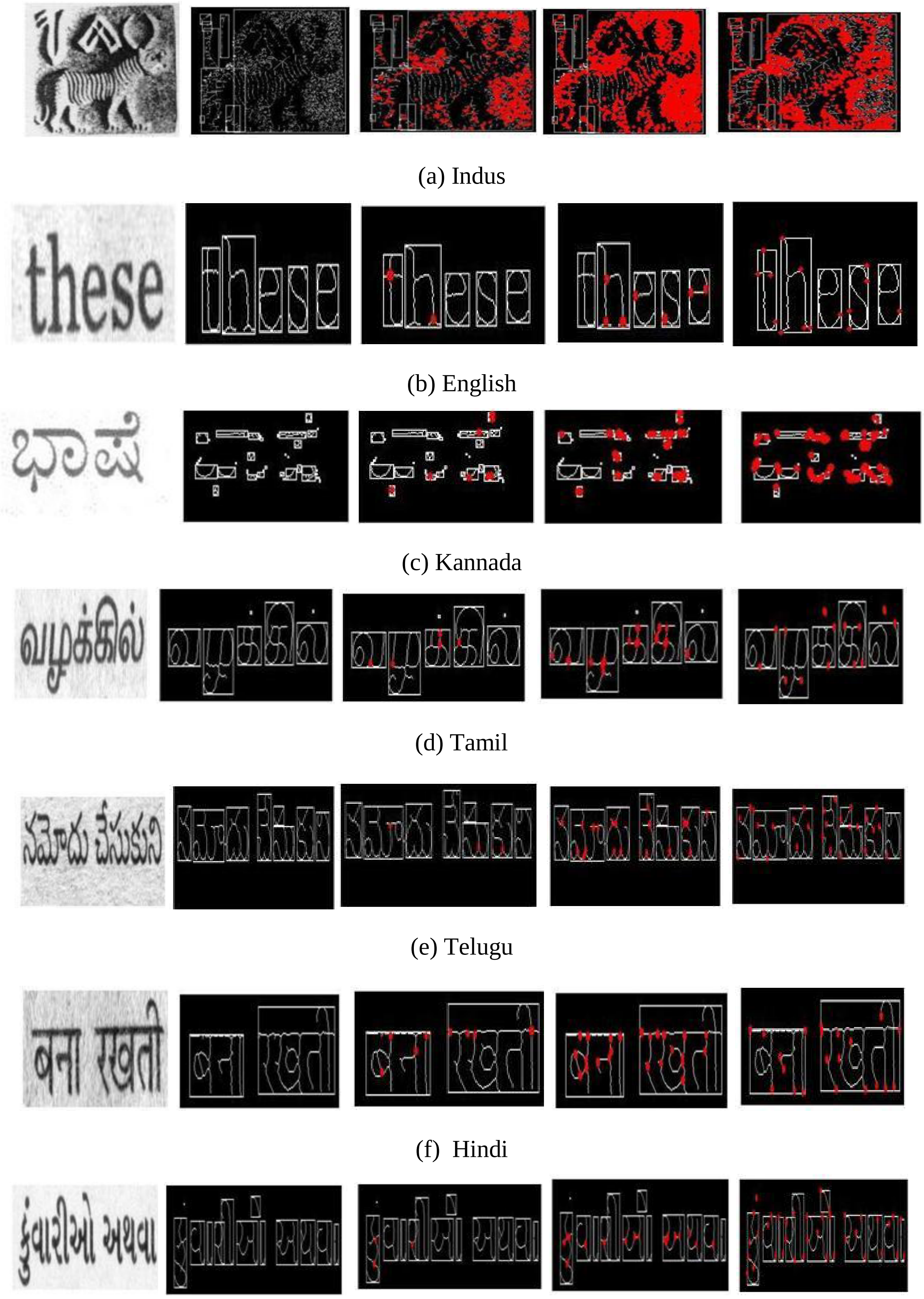
(c) Bonding boxes for the components in text line image

(d) Intersection points (marked by red color)

(e) Junction points (marked by red color)

(f) End points (marked by red color)

Fig. 3.17.Intermediate steps for the dominant pixel extraction



(a) Indus

(b) English

(c) Kannada

(d) Tamil

(e) Telugu

(f) Hindi

(g) Gujarati

Fig. 3.18. Three features with respect to three dominant pixels for seven scripts

3.2.3.2 New Variance based Features

We can see more number of Intersections, Junction and End points in Indus compare to English and other scripts. In English it is found less number of intersection points due to fewer strokes. In some scripts like Tamil joints of strokes are found in middle. In Telugu dominant points found more in middle and bottom portion. However position of dominant pixels in the text script varies due to structure of scripts. Hence it is necessary to identify relationship between dominant pixels. Thus we measure the distance between dominant pixels to study the spatial relationship among dominant pixels.

The proposed method computes the distance matrix for each set of dominant points. Distance between each of the corresponding dominant pixel is computed resulting in three different distance matrix one for each of the dominant pixels. Then, the method computes variance for these matrices. Variance is found for each component. Then the method finds the variances for whole text line. Variance of text line is the sum of variance of each components divided by number of components. Similarly, variance detected fortext line for each set of dominant pixels. For example let $D = \{d1, d2, \dots, dn\}$ be the dominant pixels of certain type.Distance from d1 is computed to all the points$\{d1, \dots, dn\}$ then ,distance from d2 is computed to all points $\{d1, \dots, dn\}$ .Similarly distance from all the points are computed forming distance matrix. Distance between the two dominant pixels is measured using Euclidean distance.

$$\mathrm{Euclid}_{\mathrm{dist(x1,x2)}} = \mathrm{SQRT}((\mathrm{x2} - \mathrm{x1})^2 + (\mathrm{y2} - \mathrm{y1})^2) \qquad (3.18)$$

Let intr, junc, endp are the array of intersection, junction and endpoints respectively.Distance matrix computed is$IDM(i,j), JDM(i,j), EDM(i,j)$for intersection, junction and End points where iis the number of rows and j is the number of columns.

$$\mathrm{IDM(i,j)} = \forall_{i,j=1,\ldots,n}\mathrm{Euclid}_{\mathrm{dist}(\mathrm{intr_i},\mathrm{intr_j})} \tag{3.19}$$

$$\mathrm{JDM(i,j)} = \forall_{i,j=1,\ldots,n}\mathrm{Euclid}_{\mathrm{dist}(\mathrm{junc_i},\mathrm{junc_j})} \tag{3.20}$$

$$\mathrm{EDM(i,j)} = \forall_{i,j=1,\ldots,n}\mathrm{Euclid}_{\mathrm{dist}(\mathrm{endp_i},\mathrm{endp_j})} \tag{3.21}$$

For each component distance matrix are computed as in Equation (3.19-3.21).Let $VARI, VARJ, VARE$ be the variance computed for intersection, junction and end points respectively. Variance $VARI, VARJ, VARE$for each of the component is computed as in Equation (3.22-3.24).

$$VARI = VAR\big(IDM(i,j)\big) \tag{3.22}$$

$$VARJ = VAR\big(JDM(i,j)\big) \tag{3.23}$$

$$VARE = VAR\big(EDM(i,j)\big) \tag{3.24}$$

Let $AVGI, AVGJ, AVGE$ are the average of variances of line corresponding to Intersection, Junction, End points .Average of variance is computed as in Equation (3.25-3.27).

$$AVGI = \frac{1}{N}\sum_{i=1}^{N} VARI_i \tag{3.25}$$

$$AVGJ = \frac{1}{N}\sum_{i=1}^{N} VARI_j \tag{3.26}$$

$$AVGE = \frac{1}{N}\sum_{i=1}^{N} VARI_e \tag{3.27}$$

Let $F1, F2, F3$ denote the features for intersection points, junction points and end points, respectively. These three features correspond to values computed by $AVGI, AVGJ, AVGE$.Thus $F1, F2, F3$ are the variance computed for the line withintersection, junction, and end points respectively. Thus $F1, F2, F3$ are the values corresponding to the dominant pixel distribution. Such values are computed for all seven different scripts. A discriminative change of values computed as features for seven different scripts of Fig.3.18 is shown in Fig.3.19.

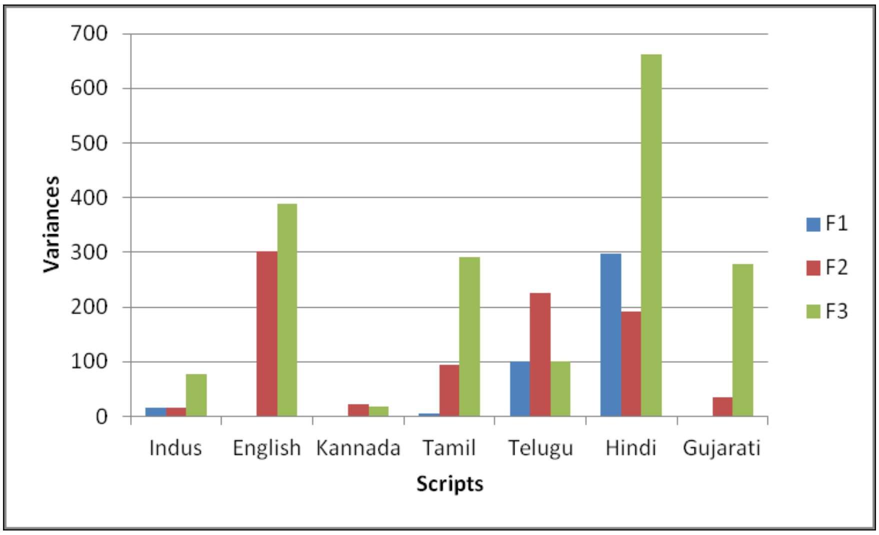


Fig. 3.19. Three features dominant pixels for seven scripts

3.2.3.3 Script Identification through Template Construction

Templates are the standard Feature values computedfor further identification of scripts. Feature values in Fig.3.19 shows clearly the structural difference among scripts. Hence for automatic script identification,templates are generated. To generate templates we choose 50 samples randomlyfrom each kind of script. Variance are computed for each of these 50 samples and we compute the average of those feature values to compute templates for feature values.In total we chose 350 training samples to generate templates.Let $TF1, TF2, TF3$ are the feature values computed using Equation (3.28 – 3.30).

$$TF1 = \frac{1}{50}\sum_{i=1}^{50} F1(i) \tag{3.28}$$

$$TF2 = \frac{1}{50}\sum_{i=1}^{50} F2(i) \tag{3.29}$$

$$TF3 = \frac{1}{50}\sum_{i=1}^{50} F3(i) \tag{3.30}$$

In this way, template values are generated for all seven different scripts. For each of the script types $TF1, TF2, TF3$ are found that correspond to dominant points namely, Intersection, Junction and End points. Therefore seven templates generated for each dominant points.Template values distribution for seven different scripts is as shown in Fig 3.20. To identify script, Feature values $F1, F2, F3$ are computed for each of the connected component using Equation (3.22-3.24). Then average of the feature values is predicted using Equation (3.25-3.27). Then the method compares the feature values of the script under test with the seven template values generated by the corresponding dominant point. Method computes minimum Euclidean distance for comparison. Template that gives the minimum distance determines the script type.

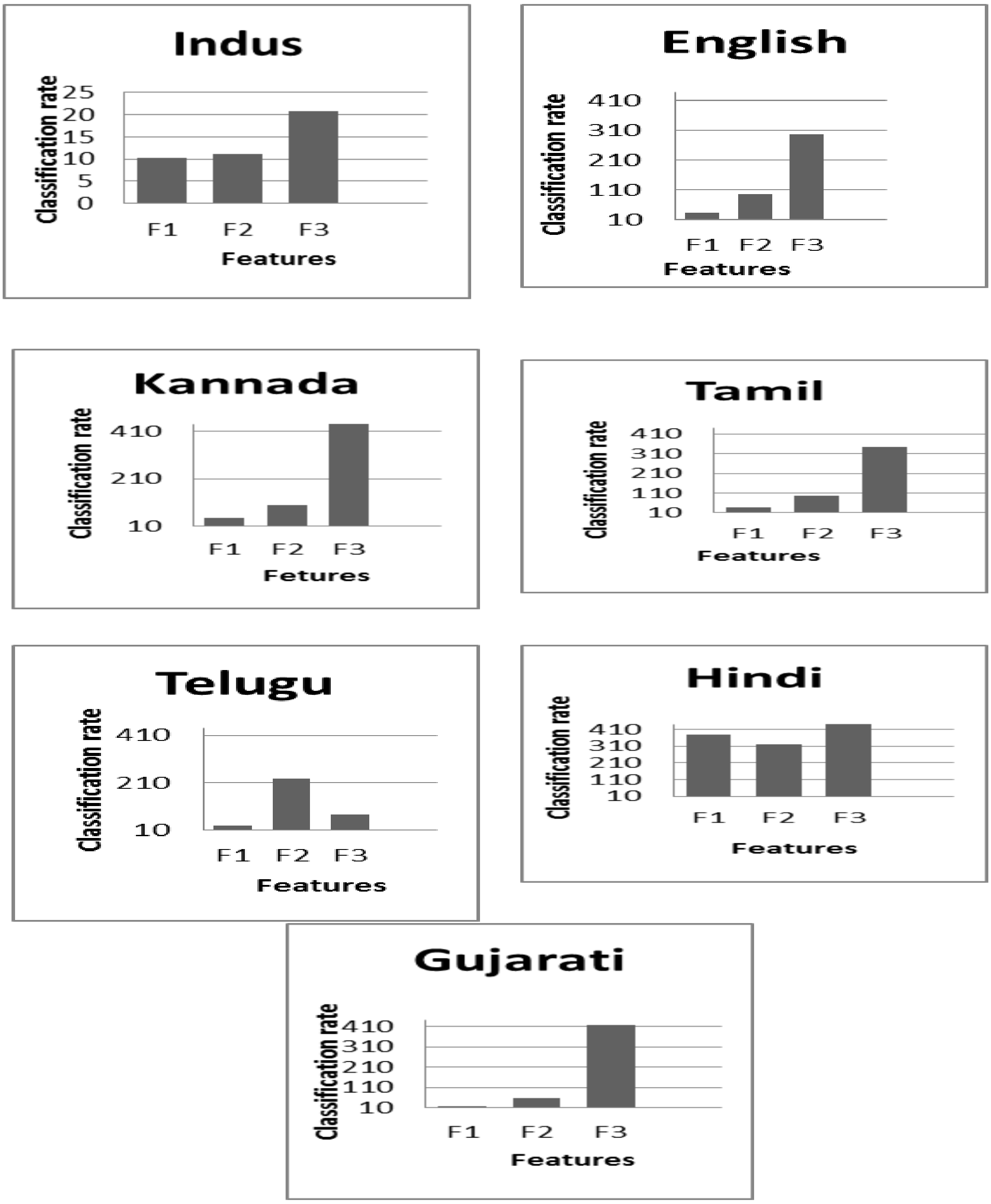


Fig. 3.20. Templates generated for different scripts

3.3 Experimental Results

To conduct experiments on the various approaches we collect document images from various sources like newspapers, magazines, books. Wescan 800 document images (100 Indus, 100 English, 100 Kannada, 100 Tamil, 100Telugu, 100Malayalam, 100 Hindi and 100 Gujarati documents). Particularly, Indus documents are collected from an Archeological Survey of India (ASI), Mysore in Karnataka state of India. Since there is no benchmark/standard database for Indus and documents of different languages, we create our own data base for experimentation and evaluation.

3.3.1 Experiments for Centroid based Straightness and Cursiveness Approach

We conduct two different experiments choosing 600 documents (100 Indus, 100 English, 100 Kannada, 100 Tamil, 100 Telugu, 100Malayalam) from the dataset to evaluate the performance of the method. Methods are based on centroid to study Straightness and Cursiveness of the components. In first experiment Indus documents are identified from other documents. In second experimentEnglish documents are identified from other documents. To evaluate the effectiveness of the proposed method we experiment our dataset on existing method Padma and Vijay 2009. Reason for choosing this existing method is that, this method focuses on English and some of the South Indian scripts.

We report quantitative results of the proposed centroid based, region growing based and integrated methods, and existing method in Table 3.1 and Table 3.3, respectively. The centroid based method is tested on Indus and other documents classification. Here other document includes four south Indian and English documents. Similarly, the region growing based method is tested on English and South Indian documents. Finally, we combine both the methods to classify Indus, English and South Indian documents, which we called proposed integrated method. Table 3.1 and Table 3.2 and Table 3.3 show that the proposed methods score promising results for classification of Indus from other documents and English from south Indian documents. However, the proposed integrated method classifies Indus and English with good classification rate, at the same time, it fails to classify the south Indian scripts. This is due to poor features.

Table 3.1: Confusion matrix for Proposed Centroid based Method

Proposed Method		
Scripts	Indus	Others
Indus	**67 %**	33 %
Others	35.8%	**64.2%**

Table 3.2:Confusion matrix for proposed Region based Method

Proposed Method		
Scripts	English	Others
English	**80%**	20%
Others	34.7%	**65.2%**

Table 3.3:Confusion matrix for the Proposed Integrated Method

Integrated Method	Indus	English	South Indian Scripts
Indus	**67**	0	33
English	20	**80**	0
South Indian Scripts	95	0	**5**

To show effectiveness of the proposed integrated method, we compare with the Padma and Vijay method. Since the existing method is developed for classifying English from other documents, we conduct the experiments by considering English documents as one class and other documents as another class. The quantitative results of the proposed and existing method are reported in Table 3.4 where we can see that the proposed integrated method scores better results than existing method. The main reason for poor results of the existing method is that the method is developed for plain documents but not for the scanned document images.

Table 3.4: Comparative study of the Proposed Integrated and the Existing method

Methods	English document identification (in %)
Padma and Vijaya[ ]	34
Proposed Method	**80**

3.3.2 Experiments for Skew and Orientation based Straightness and Cursiveness Approach

Methods discussed in section 3.3.1 gives poor performance in Identifying Indus documents and also the method also results in false positives in identifying English from others. To overcome these disadvantages we develop improved methods based on skew and orientation of the components. We follow the same experimental set up used in previous section to evaluate the proposed methods. Experimental results of the region skew based method, nearest neighbor clustering based method and integrated methods are reported in Table 3.5-Table 3.7, respectively. When we compare the results of these methods with the results of the methods discussed in previous section, the results of region skew based and nearest neighbor clustering based method give better results.

Table 3.5:Confusion matrix for proposed Region Slope Skew based Method

Proposed Method (2 classes).		
Scripts	Indus	Others
Indus	**98 %**	02 %
Others	28.8 %	**71.2%**

Table 3.6: Confusion matrix for proposed Nearest Neighbor Clustering based Method

Proposed Method (2 classes).		
Scripts	English	Others
English	**83 %**	17 %
Others	50.75 %	**49.25%**

Table 3.7:Confusion matrix for the proposed integrated method

Integrated Method	Indus	English	South Indian Scripts
Indus	**98**	0	2
English	4	**90**	6
South Indian Scripts	5	5	**90**

Table 3.8 reports comparative of the integrated method discussed in previous section and the integrated method discussed in this section, and the existing method. The integrated method discussed in this section score better results compared to the other methods. The reason for the poor results of the existing method is same as mentioned in previous section.

Table 3.8:Comparative study of the proposed integrated methods and existing method

Methods	English document identification (in %)
Padma and Vijaya[ ]	44
Kavitha et al. [ ]	80
Proposed Method	**83**

3.3.3 Experiments for Spatial Relationship Based Approach

It is noted from the previous experiments that the scope of the methods is limited to three classes which includes Indus, English and South Indian scripts. In this section, we conduct experiments on spatial relationship based method which considers Indus, English, Kannada, Tamil, Telugu, Hindi and Gujarati. Different scripts included are 100 Indus, 100 English, 100 Kannada, 100 Tamil, 100 Telugu, 100 Gujarati and 100 Hindi. In total, 700 documents are considered for experimentation. We choose 50 samples from each class randomly to create template as discussed in Section 3.2.3.3. The results of the method are reported in Table 3.9. This experiment gives the best classification rate. The spatial relationship based method used Canny edges for extracting features. To show that Canny is better than other edge detectors, namely Sobel, we conduct experiments using Sobel edge as reported in Table 3.10. It is observed from Table 3.10 and Table 3.11 that Canny is better than Sobel edges. This is valid because Sobel lose many pixels when the images have low contrast while Canny is good for both low and high contrast images. Instead of choosing 50 training samples randomly from each class, we use five-fold cross validation technique for generating confusion matrix and the results are reported in Table 3.12-Table 3.16. The final average of all five-fold confusion matrix is reported in Table 3.17. When we compare Average classification rate given by five-fold cross validation and the 50 ransom samples, the results given by ransom sample is higher than the five-fold cross validation. This shows that the method loses accuracy when the number of training samples reduces. Therefore, five-fold cross validation is not good for small dataset. In this work, we divide the whole dataset into five equal groups. Each group is considers at a time for creating templates and rest of the groups are used for testing. Therefore, the number of training samples is lower than 50% which is considered for random sampling.

Table 3.9: Confusion matrix for the proposed method in %

Script type	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**98**	0	1	0	0	1	0
English	0	**97**	0	2	1	0	0
Kannada	0	0	**97**	0	0	0	3
Tamil	0	3	0	**93**	1	0	3
Telugu	0	1	0	0	**98**	0	1
Hindi	1	0	4	0	0	**95**	0
Gujarati	0	0	2	2	1	0	**95**

Table 3.10: Confusion matrix using Sobel edges.

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	83	0	0	1	0	1	0
English	0	9	1	79	2	0	9
Kannada	0	0	60	0	0	40	0
Tamil	0	6	3	48	3	0	40
Telugu	4	23	0	26	45	0	2
Hindi	4	3	17	1	2	67	6
Gujarati	0	0	78	0	0	20	2

Table 3.11: Confusion matrix using Canny edges.

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**98**	1	0	0	0	1	0
English	0	**7**	0	1	91	0	1
Kannada	0	0	**48**	6	0	0	46
Tamil	0	41	0	**34**	23	0	2
Telugu	6	2	0	2	**92**	0	0
Hindi	5	7	27	6	6	**32**	17
Gujarati	0	0	4	14	1	0	**81**

Table 3.12: Confusion Matrix of first fold

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**97**	0	1	0	1	1	0
English	0	**96**	1	2	1	0	0
Kannada	0	0	**97**	0	0	0	3
Tamil	0	5	0	**91**	1	0	3
Telugu	0	2	0	0	**98**	0	0
Hindi	1	0	4	0	0	**95**	0
Gujarati	0	0	2	3	1	0	**94**

Table 3.13: Confusion Matrix of second fold

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**96**	1	1	0	1	1	0
English	0	**97**	1	2	0	0	0
Kannada	0	0	**96**	0	0	0	4
Tamil	0	4	0	**90**	1	0	5
Telugu	0	8	0	2	**90**	0	0
Hindi	1	0	5	0	0	**94**	0
Gujarati	0	0	2	3	0	0	**95**

Table 3.14: Confusion Matrix of third fold

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**97**	0	1	0	1	1	0
English	0	**97**	0	2	1	0	0
Kannada	0	0	**97**	0	0	0	3
Tamil	0	5	0	**91**	1	0	3
Telugu	0	2	0	2	**96**	0	0
Hindi	1	0	2	0	0	**97**	0
Gujarati	0	0	2	2	0	0	**96**

Table 3.15: Confusion Matrix of fourth fold

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**98**	0	1	0	0	1	0
English	0	**95**	1	3	1	0	0
Kannada	0	0	**97**	0	0	0	3
Tamil	0	5	0	**91**	1	0	3
Telugu	0	4	0	2	**94**	0	0
Hindi	1	0	6	0	0	**93**	0
Gujarati	0	1	1	2	0	0	**96**

Table 3.16: Confusion Matrix of fifth fold

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**97**	0	1	0	1	1	0
English	0	**97**	0	2	1	0	0
Kannada	0	0	**93**	0	0	0	7
Tamil	0	3	0	**93**	1	0	3
Telugu	0	2	0	0	**98**	0	0
Hindi	1	0	6	0	0	**93**	0
Gujarati	0	1	3	1	0	0	**95**

Table 3.17: Average classification rates

	Indus	English	Kannada	Tamil	Telugu	Hindi	Gujarati
Indus	**97**	0.2	1	0	0.8	1	0
English	0	**96.4**	0.6	2.2	0.8	0	0
Kannada	0	0	**96**	0	0	0	4
Tamil	0	4.4	0	**91.2**	1	0	3.4
Telugu	0	3.6	0	1.2	**95.2**	0	0
Hindi	1	0	4.6	0	0	**94.4**	0
Gujarati	0	0.4	2	2.2	0.2	0	**95.2**

3.3.4 Comparative Study

In this section, we implement three existing methods (Padama and vijay, Das et al and Hangargeaet al) to give comparative study with the proposed integrated method-1, proposed integrated method-2 and the spatial relationship based method. The results are recorded in Table 3.18 where we can see the proposed methods are better than existing methods. This is due to the existing methods are not capable of handling complex documents like Indus. In addition, the existing methods are developed for plain background images but not scanned images. Note that since the existing methods are developed for classifying English documents from other documents, we consider English documents as one class and other documents as one class.

Table 3.18: Comparative study table for existing methods and all proposed methods.

Methods	Identification rate in %
Padma &Vijaya, 2010	7
Das et al., 2011	47
Hangargea et al., 2013	44
Proposed method-Integrated method-1 (Table 3.3)	**80**
Proposed method-Integrated method-2(Table 3.7)	**90**
Proposed Method- Spatial Relationship of Dominant Points(Table 3.15)	**96.1**

3.4 Summary

In this chapter, we have proposed centroid based, region skew based and spatial relationship based methods for classifying Indus scripts, English and other Indian scripts. The methods explored centroid of the components, direction of the components, and spatial relationship between the pixels for extracting distinct features for identifying scripts. The proposed integrated method-1 works well for two classes, the proposed integrated method-2 works well for three classes and the proposed spatial relationship based method works well for seven classes. Experimental results show that the proposed method is better than existing methods in terms of classification rate. Our future work would be expanding the proposed method for other Indian and Foreign scripts.

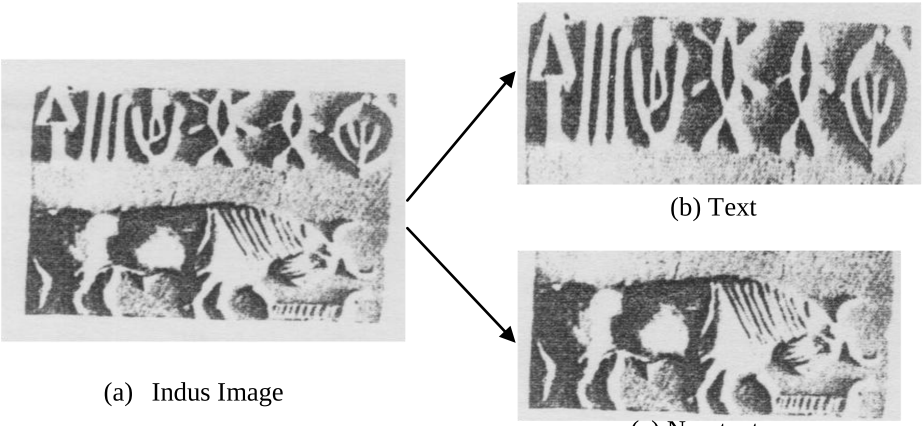
Chapter 4

Text line segmentation

4.1 Background

In chapter 3, we have proposed method for script identification to facilitate text line segmentation because each script has its own nature and characteristics. As a result, it is hard to develop unified segmentation method for different scripts.

In this chapter, we propose a method for text line segmentation. This chapter focus on text line segmentation for Indus document. It is noted from the Indus document that text components have less cursive compared to non-text components. Based on this, the proposed method segment text lines from Indus document images. Sample text line segmentation is shown in Fig. 4.1 where we can see text different symbols for text and animal like picture for non-text components.



(a) Indus Image

(b) Text

(c) Non text

Fig. 4.1. Illustrating text and non-text components in Indus documents

Some parts of the material of this chapter appeared in the following research papers:

1. "Text Line Segmentation in Degraded Historical Documents",Egyptian Informatics Journal,Vol 17(2),July 2016, pp 189-197.

4.2 Proposed Methodology

Since Indus document suffer from low contrast, we propose the combination of sobel and Laplacian edge images for enhancement of low contrast pixels. Then the proposed method extract skeleton to reduce pixel width of edge components to single pixel width which help us to study the structure of text components. Then the components are merged based on nearest neighbor to group the components. The process of merging continues until it gives two clusters. Braches of each cluster components has been studied to classify text and non-text. Thus the cluster that has less number of branches is proved to be text part while cluster that has more number of branches is the non text part in Indus images. The flow of the proposed methods is shown in Fig. 4.2.

The proposed methodology is divided into three subsections. In first subsection, we explore the combination of sobel and Laplacian edge images to enhance the low contrast pixels. In second subsection, we introduce the nearest neighbor criterion for grouping nearest components. In third subsection, we discuss the cluster formation for segmentation of text and non text in Indus document images.

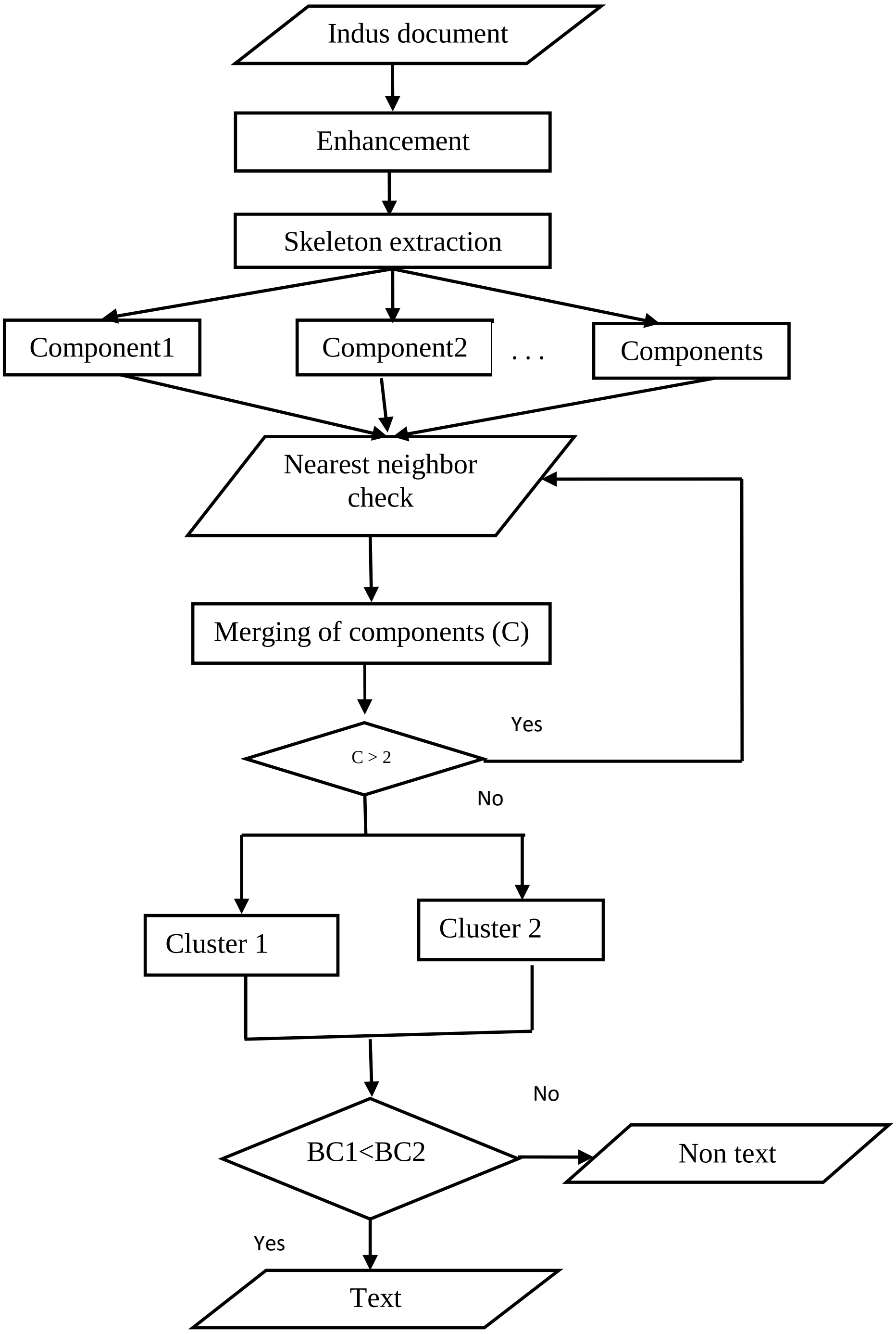


Fig. 4.2. Flow diagram of the proposed text line segmentation: BC1 and BC 2 denote the number of branches in cluster-1 and cluster-2, respectively.

4.2.1 Text Enhancement

This section proposes method for enhancement of Indus document images. It is noted that Sobel gives fine edges for high contrast pixels and Laplacian gives fine edges for both low and high contrast pixels. We exploit this observation for enhancing low contrast text pixels in Indus image. Therefore, we perform intersection operation which results in common pixels which represent text information as shown in Fig. 4.3 where we can see the intersection operation eliminates most of background pixels. In order to study the structure of text components, the proposed method obtains skeleton for the intersection results as shown in Fig. 4.3(d). Mathematically, steps of the enhancement are represented as follows.

The masks to compute Sobel gradient Image are given by,

$$Gx = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad And \quad Gy = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \qquad (\mathbf{4.1})$$

The Sobel gradient of the image for array G [i, j] is obtained by finding the gradient Gx in x direction and Gy in y direction.

$$Gx = ((2 * \mathrm{C}(i+2, j+1) + \mathrm{C}(i+2, j) + \mathrm{C}(i+2, j+2)) - (2 * \mathrm{C}(i, j+1) + \mathrm{C}(i, j) + \mathrm{C}(i, j+2))) \qquad (\mathbf{4.2})$$

And

$$Gy = ((2 * \mathrm{C}(i+1, j+2) + \mathrm{C}(i, j+2) + \mathrm{C}(i+2, j+2)) - (2 * \mathrm{C}(i+1, j) + \mathrm{C}(i, j) + \mathrm{C}(i+2, j))) \qquad (\mathbf{4.3})$$

The magnitude of pixel gradient is given by,

$$\mathrm{G}[i, j] = \left|\sqrt{Gx^2 + Gy^2}\right| \qquad (\mathbf{4.4})$$

Laplacian Gradient for array C [i, j] is computed as

$$Kx = \left(\left(\mathrm{C}(\mathrm{i}+2,\mathrm{j}+1) + \mathrm{C}(\mathrm{i},\mathrm{j}+1) - 2 * \mathrm{C}(\mathrm{i},\mathrm{j})\right)\right) \qquad (\mathbf{4.5})$$

$$Ky = \left(\left(\mathrm{C}(\mathrm{i}+1,\mathrm{j}+2) + \mathrm{C}(\mathrm{i}+1,\mathrm{j}) - 2 * \mathrm{C}(\mathrm{i},\mathrm{j})\right)\right) \qquad (\mathbf{4.6})$$

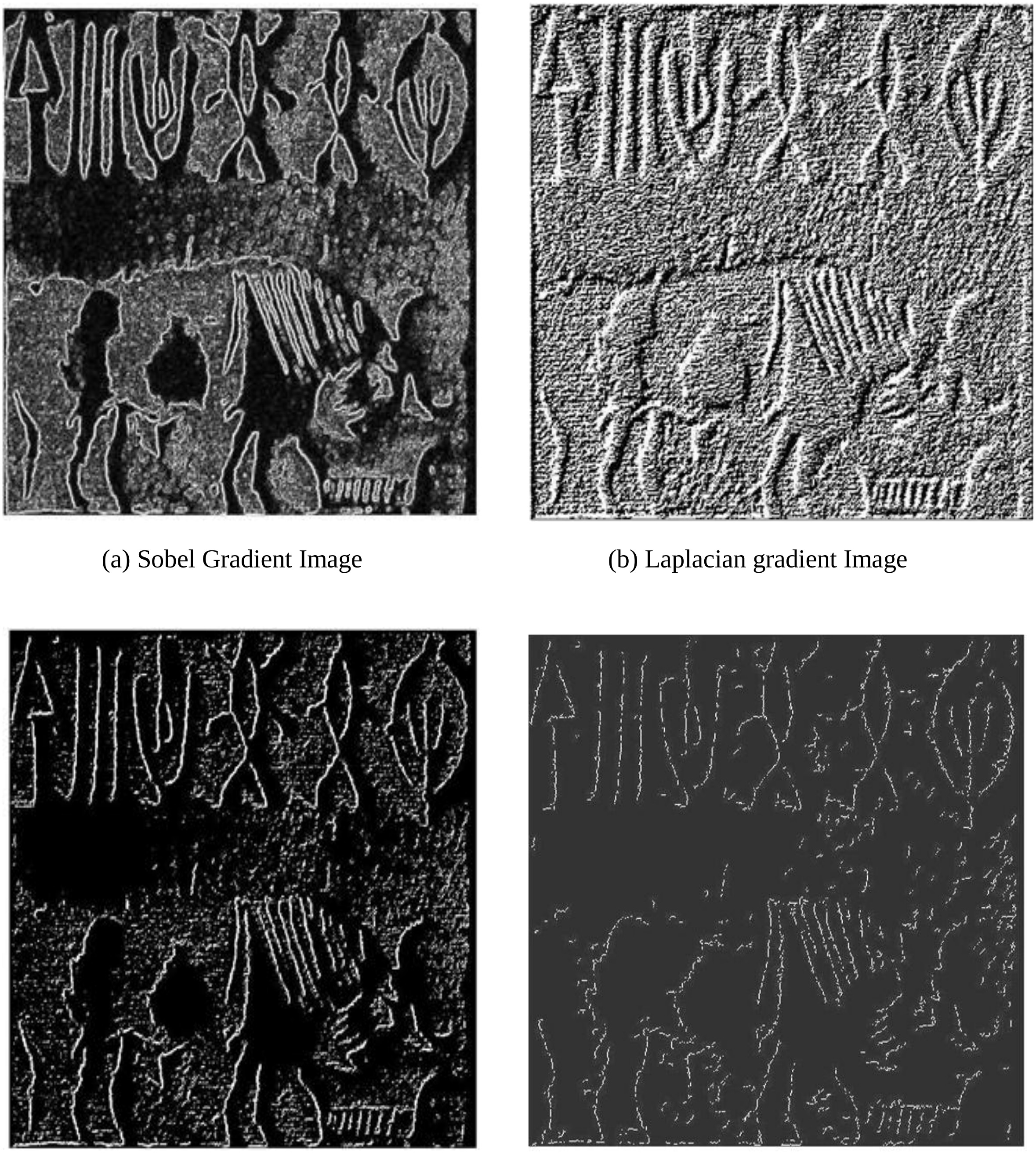
$$K[I,J] = Kx + Ky \qquad (\mathbf{4.7})$$

The intersection of gradient images is obtained by

$$G[I,J] \cap K[I,J] \qquad (\mathbf{4.8})$$

$$Area = Length * Height \qquad (\mathbf{4.9})$$

Skeleton obtained from text enhancement as in Fig. 4.3 (d) contains enhanced text and non text components. As seen there also exist some noisy speckles other than text and non text components. To remove noisy speckles area of each componentfound with the positional coordinates of bounding box. Bounding box with smaller area is removed by calculating using Equation 4.9 and the smoothed image obtained as in Fig. 4.4(a).



(a) Sobel Gradient Image

(b) Laplacian gradient Image

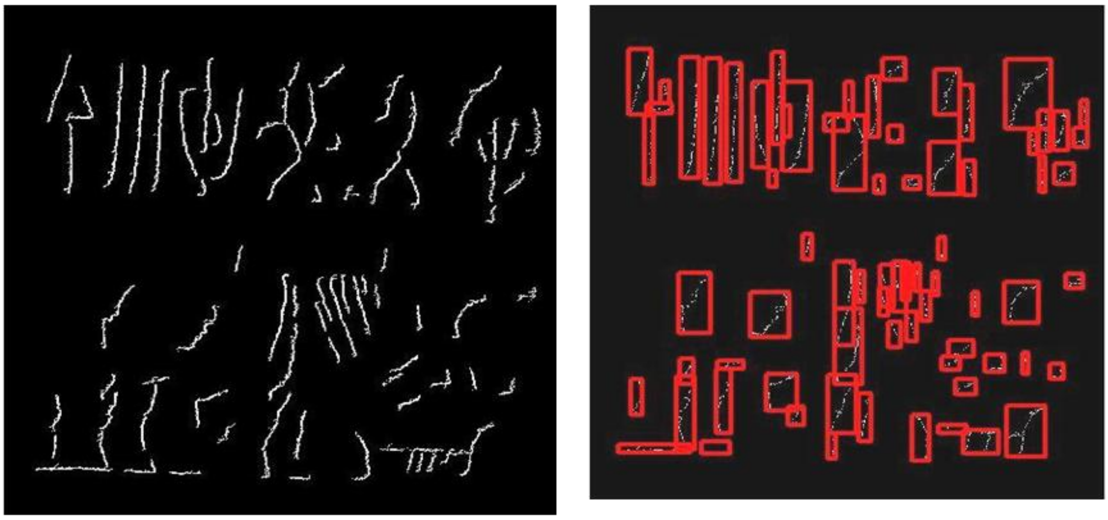
(c) Intersection of two gradient Images

(d) Skeleton

Fig. 4.3. Intermediate results of text enhancement method

4.2.2 Components Grouping

It can be seen from Fig 4.4(a) that all the texts symbols aligned horizontally with regular spacing while non-text which looks like animal does not. For each component in Fig. 4.4(a), the proposed method fixes bounding box as shown in Fig. 4.4(b). Since the space between text component is lesser than the space between text lines, the proposed method use nearest neighbor criteria for merging components by searching horizontally and vertically as shown in Fig. 4.5. The proposed method merges the components that have overlapping bounding boxes horizontally and vertically. In this way, the proposed method group text and non-text components separately. However, to determine which group represent actual text components, the proposed method considers the number of branches because if the group contains the components like animal picture, one can expect more branches compared to group which contains text components.

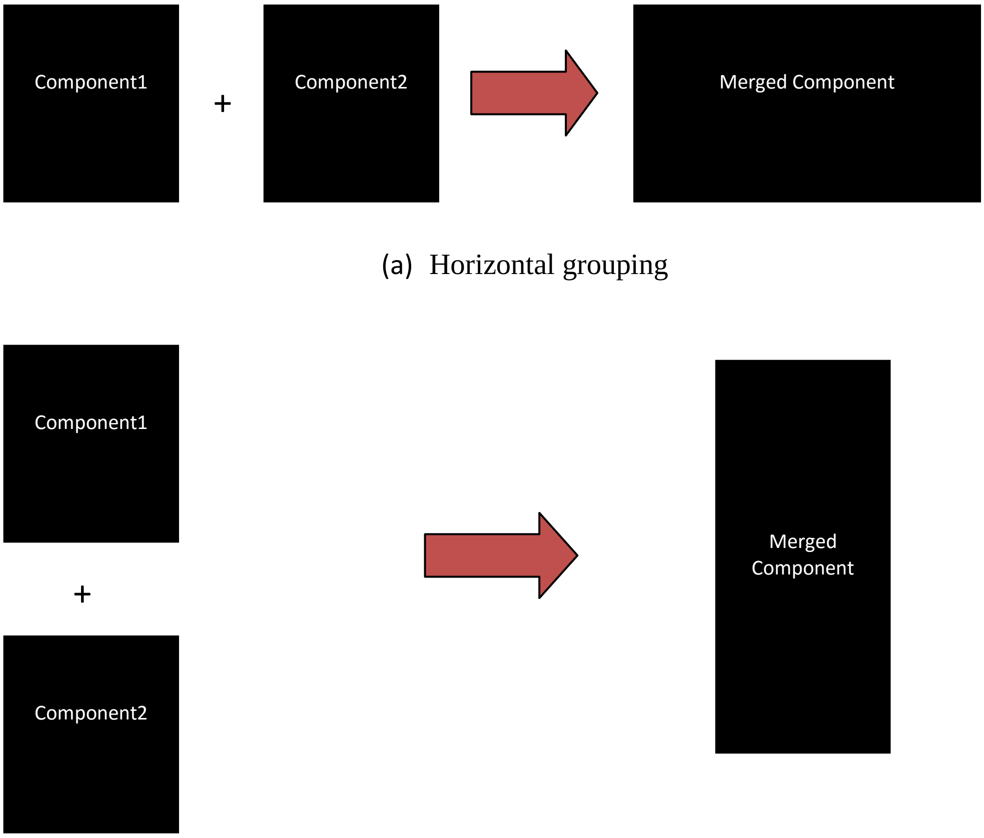


(a) Smoothed image

(b) Components with bounding box

Fig. 4.4. Illustration of grouping of pixels.

.



(a) Horizontal grouping

(b) Vertical grouping

Fig 4.5. Illustration of grouping of components

4.2.3 Cluster formation for Text Line Segmentation

For each component in the image, the method finds the nearest neighbor component using Euclidean distance as defined in Equation 4.10. The component that gives the minimum distance is considered for grouping. Series of steps involved in formation of clusters is shown in Fig 4.8. Cluster formation for horizontal grouping is as shown in Fig 4.6(a-e). Horizontal grouping resulted in three clusters. As per the proposed method, grouping of the components should result in 2 clusters to define them as text and non text. Hence the proposed method groups the horizontal clusters using vertical grouping as in Fig 4.5(b). Hence process of grouping of clusters continues until the resultant cluster is 2. As seen in Fig 4.6 (f) grouping of clusters results in two clusters. Process of cluster formation can be illustrated mathematically as follows. Let $\mathrm{C} = \{\mathrm{C1}, \mathrm{C2}, \mathrm{C3}, \ldots, \mathrm{Cn}\}$ be the finite set of components. Let C_i be candidate components considered for merging. The Distance between a candidate component and another component is calculated using the boundary values of the components. C_j

is the set of all the other components excluding $\mathrm{C_i}$. The distance between the two components is considered for merging of components is determined as follows.

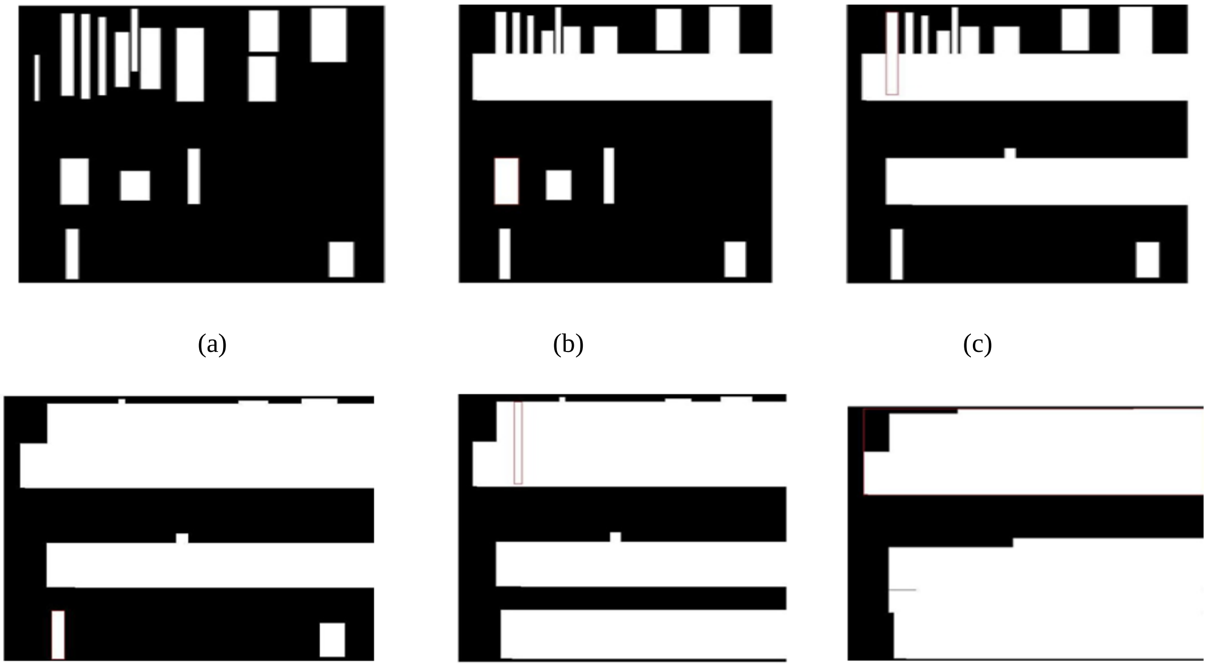
Let$(X_1, Y_1) and\ (X_2, Y_2)$ be the extreme coordinates of two neighboringcomponents. Distance between the two components is obtained by,

$$\mathrm{Euclid\_dist}_{\mathrm{x1,x2}} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \qquad (\mathbf{4.10})$$

The proximity of closeness defined by the minimum distance criteria is calculated by,

$$\mathrm{Cn} = \mathrm{Min}\ \{\mathrm{d(a, b)}: \mathrm{a\epsilon C_i}, \mathrm{b\epsilon C_{j=1,n}} - \mathrm{C_i}\} \qquad (\mathbf{4.11})$$

Cluster formation results in two clusters without identifying cluster as text andnon text. Therefore, we propose branch feature to extract number of branches from clusters. We find number of branches from the skeleton in the boundary of 2 clusters to identify text and non text clutters. The proposed method compares the number of branch points from the corresponding skeleton of each clustered image.

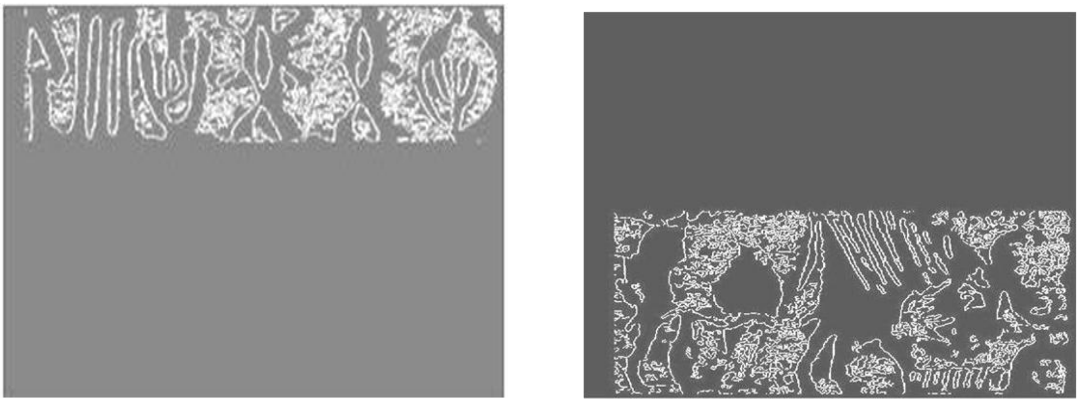


(a) (b) (c)

(d) (e) (f)

Fig. 4.6. Illustration of grouping process for Fig.4.4 (b)

The cluster that gives less number of branches is considered as a text cluster. The cluster that gives a more number of branches is considered as a non text cluster. This is because text components usually have fewer branches compared to non-texts due to presence of animal like picture. Let NB_{c1} and NB_{c2} be the numbers of the branches in cluster1 and cluster2, respectively. If $NB_{c1} < NB_{c2}$then the contour of clusters c1 is identified as text and cluster c2 is identified as non text. If $NB_{c2} < NB_{c1}$ then the contours of clusters c2 is identified as text and cluster c1 is identified as non- text. The results obtained after such comparison for sampled image of Fig. 4.1(a) isidentified as text cluster as in Fig. 4.7(a) and as non- text cluster as in Fig. 4.7(b).



(a) Text cluster (b) Nontext cluster

Fig. 4.7. Text and non-text cluster separation

4.3 Experimental Results

To evaluate the proposed method, we use dataset consisting of 500 Indus document images. This dataset includes images of various levels of degradation. The dataset includes scripts drawn using different tools by different people. To measure the performance of the proposed method, we count the number of cluster that represents text and the cluster that represents non-text as reported in Table 4.1, where it is noted that the proposed gives a good segmentation rate for text and non-text classification.

Table 4.1.Matching matrix of the proposed method on the classification of text lines

Proposed Method	Text	Non-text
Text	**91%**	9%
Non-text	13%	**87%**

For evaluating text line segmentation, we use well-known measures such as recall and precision as in Equation 4.12 and 4.13. The definitions for recall and precision are as follows. Let G_t be the total number of the text lines counted in 500 images, T_p be the number of the text lines segmented from the proposed method, F_n be the number of the lines that are not segmented, F_p be the number of non-texts that are classified as texts.

$$Recall = \frac{Tp}{(Tp + Fn)} \quad \mathbf{(4.12)}$$

And

$$Precision = \frac{Tp}{(Tp + Fp)} \quad \mathbf{(4.13)}$$

Further to show the effectiveness of the proposed method, we implement two existing methods on text line segmentation for comparative studies. Precision measures shown in Table 4.2 also shows good performance of 97% compare to existing methods.

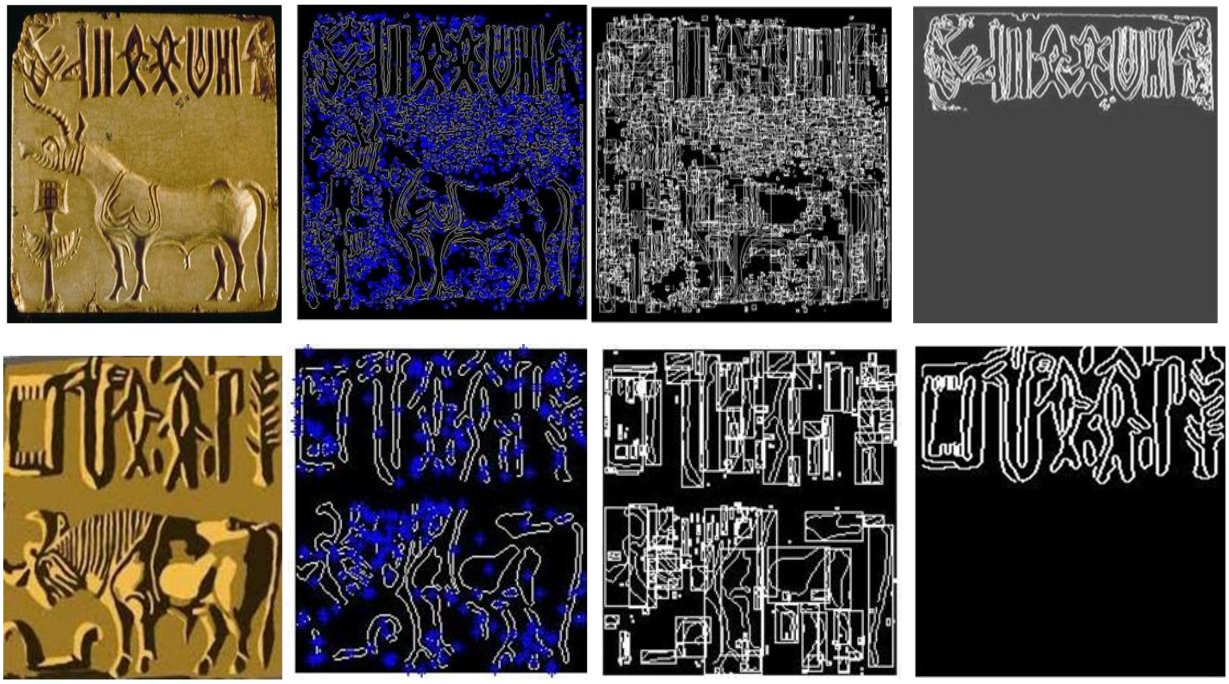
Table 4.2. Precision measures of the proposed and existing method

Method	G_t	T_p	F_n	F_p	Recall	Precision
Markus Diem et al.	515	20	495	500	0.038	0.03
Bukhari et al.	515	65	450	500	0.126	0.11
ProposedMethod	515	461	54	13	0.89	0.97

Table 4.3.Comparative study table

Methods	Text segmentation (in %)
Markus Diem et al.[ ]	4%
Bukhari et al.[ ]	13%
Proposed Method	**91%**

Method proposed by Markus Diem et al. Segments text lines based on the distance between the bounding boxes of the components in the image. The method in Bukhari et al. segmentsthe documents having text lines by drawing snakes (curves) over ridges(central points of each line). These method assumes a uniform height for all text lines. However, this is not true for Indus documents. The reason to choose these two methods is that they segment text lines irrespective of scripts, datasets and are said to be robust to non-structure layouts as in Indus documents. Since these two existing methods are developed for segmenting text lines from plain background images, the existing methods report poor accuracies for our Indus documents. The qualitative results of the proposed and the existing methods are shown in Fig. 4.8, where one can notice that the proposed method is able to segment text lines correctly for both the images,while both the existing methods fail to segment for the first input image due to the limitations of the existing methods such as the requirements of both high resolution and plain background. For the second input image shown in Fig. 4.8, the existing methods segment text and non-text lines correctly since the image contains enough spaces between text and non-text lines. Since the existing methods aims at segmenting text lines, they focus on the segmentation of text and non-text lines without separating text and non-text lines. The quantitative results of the proposed and existing methods are reported in Table 4.2 in terms of precision measures, where we can see that the proposed method shows the best values for recall and precision compared to the existing methods. The quantitative results of the proposed and existing methods in terms of percentage of text segmentation are reported in Table 4.3, proposed methods shows the best valuse compare to existing method.



(a) Input (b) Bukhari et al. (c) Diem et al. (d) Proposed method

Fig. 4.8.Text line segmentation by the existing and the proposed methods for 2 samples of Indus document images

4.4 Summary

In this chapter, we have proposed a method for segmentation of text from non text from degraded historical Indus document images. The proposed method introduces a new combination of Laplacian and Sobel operations for enhancing low contrast pixels in the images. The characteristics of the components in the image are studied to eliminate unwanted components, which results in text components pruning in the image. We have proposed a grouping process, which involves the nearest neighbor criterion for merging text components. The iterative clustering process is then proposed to separate text and non-text regions. Our future plan would be character segmentation from segmented text lines.

Chapter 5

Character segmentation

5.1 Background

In Chapter 4, we have proposed method for segmentation of text lines from Indus documents. To recognize the character, it is necessary to segment character from text lines. This chapter focuses on character segmentation by introducing watershed model. This model works based on fact that when there is a space between characters, water flows in linear passion and when there is a touching between the characters, water collects as basin.

5.2 Proposed Methodology

Proposed method considers text lines extracted from Indus document images. We use the same combination of the Sobel and the Laplacian as discussed in previous chapter at text line level rather than at image level to enhance the edge. Thus we apply morphological operations to obtain smoothed image. The smoothed image is given as input for watershed model for character segmentation. The flow of the proposed method for character segmentation in Indus document images is shown in Fig. 5.1.

The proposed methodology is divided into two subsections. In first subsection, we discuss about the enhancement of the text lines. In second subsection, we introduce watershed model for segmentation of characters.

Some parts of the material of this chapter appeared in the following research papers:

1. "A New Watershed Model based System for Character Segmentation in Degraded Text Lines", International Journal of Electronics and Communications(AEU),Vol 71,2017, 45-52.

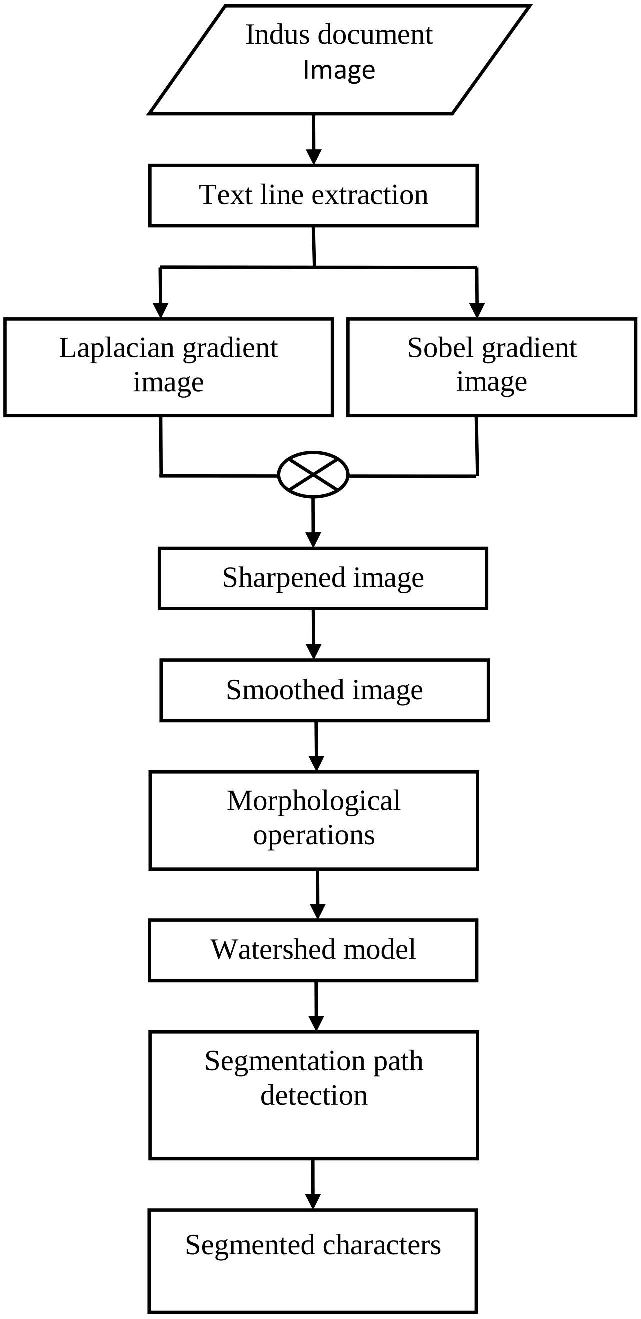


Fig. 5.1. Proposed character segmentation

5.2.1 Character Enhancement

For enhancement, text lines are extracted using the procedure discussed in chapter 4. It is a fact that Sobel gradient responds to high contrast pixels as shown in Fig. 5.3(a) resulting in thick double and sharp edges. Sobel gradient, computes first derivative measurement of an image $\nabla I(x,y)$ finds gradient in x and y direction as shown in equation 5.1.

$$\nabla I(x,y) = \sqrt{(G_X + G_y)} \tag{5.1}$$

$$\approx \nabla I(x,y) = |G_X| + |G_Y| \tag{5.2}$$

Laplacian of an image $\nabla^2 I(x,y)$computes a second derivative measurement shown in equation 5.3, is sensitive to noise as seen in Fig. 5.3(b). In order to obtain pixels which provide edges of text components, we convolve the gradients $G_X \, and \, G_Y$ images as discussed in section 4.2.1 of chapter 4. It is found convolution of Laplacian with Sobel gradient image gives the approximated smoothed image.

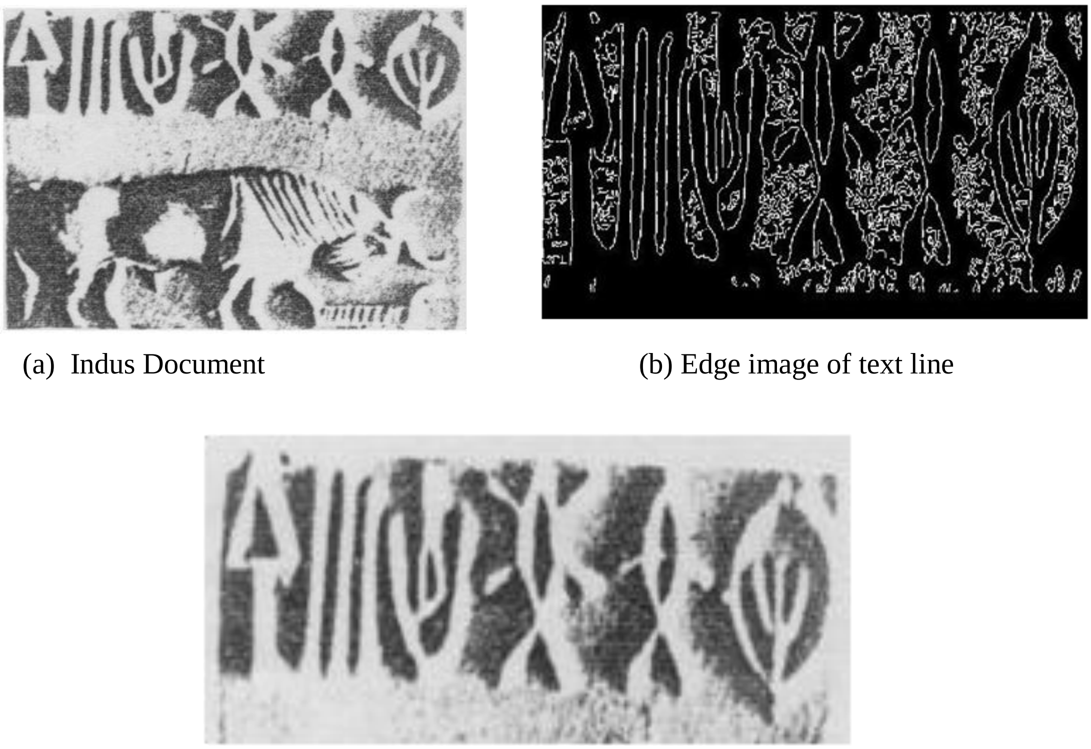
Laplacian of Image with pixel intensities $I(x,y)$ is given by,

$$\nabla^2 I(x,y) = \frac{\delta^2 I}{\delta x^2} + \frac{\delta^2 I}{\delta y^2} \tag{5.3}$$

$$\nabla^2 I(x,y) \approx 0 \tag{5.4}$$

Convolution $C(I,J)$ is performed by sliding the Laplacian image over the Sobel image for finding the intersection of these two images as in equation 5.5.

$$C(I,J) = \nabla I(x,y) \circ \nabla^2 I(x,y) \tag{5.5}$$



(a) Indus Document

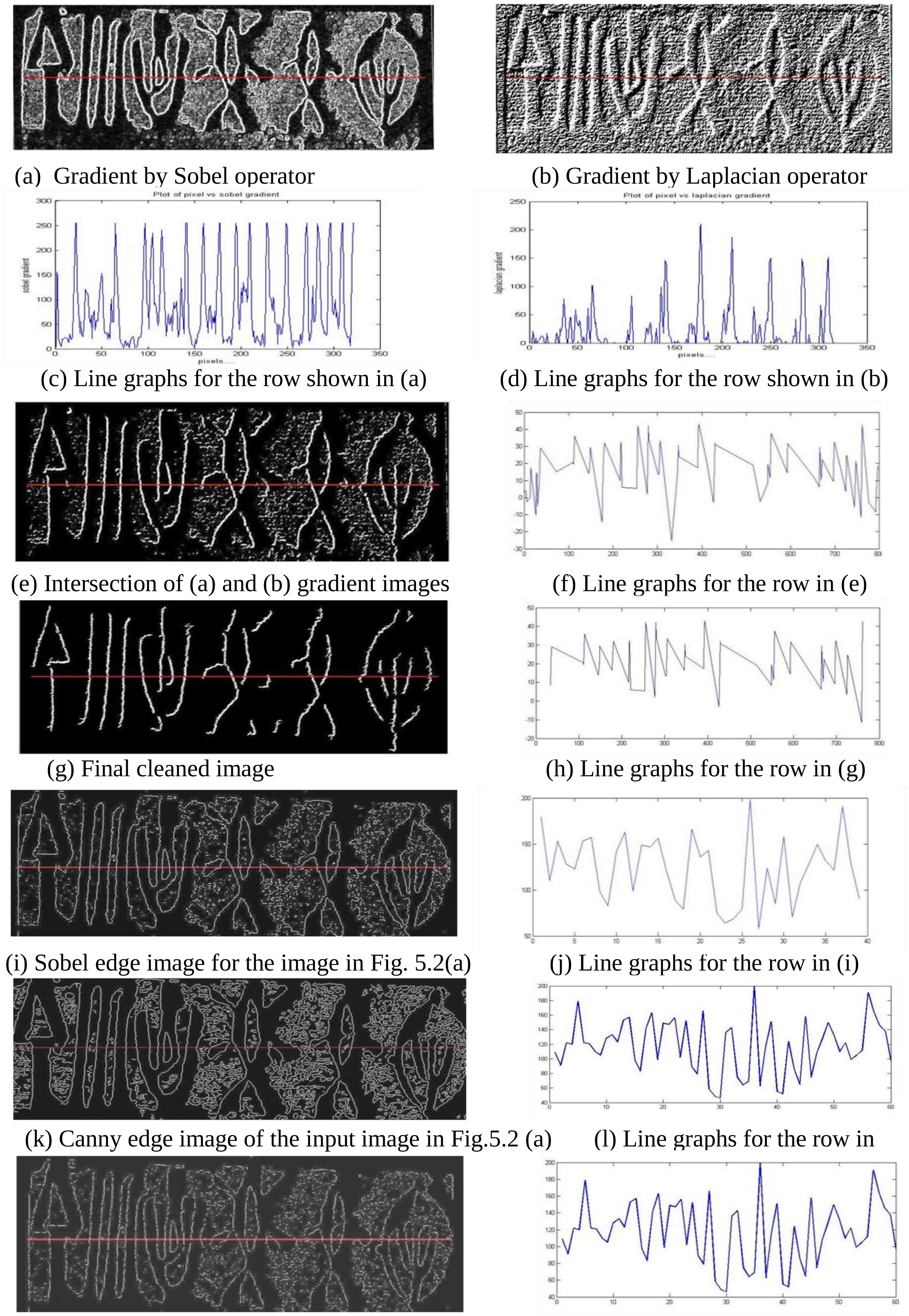
(b) Edge image of text line

(c) Text line

Fig. 5.2. Text line segmentation from Indus document image

When we look at the line graphs Fig. 5.3(c) drawn for the results in Fig. 5.3(a) it is found more number of valleys than the number of characters. Similarly, when we look at the line graphs Fig. 5.3(d) drawn for the results in Fig. 5.3(b) it has more number of peaks than the number of characters. Lines to peaks are more cursive. This shows intensity changes are not even and there are more noisy pixels. When we look at line graphs Fig. 5.3(f), drawn for the results in Fig. 5.3(e), better sharp peaks at edges and low peaks at spaces. Lines are not cursive compared to the line graphs in Fig. 5.3(c) and Fig. 5.3(d). However, Fig. 5.3(e) still contains a few background noises. To remove such small noise pixels, the proposed method uses connected component analysis to remove small components for the output of enhancement step as shown in Fig. 5.3 (g), where noise pixels are removed and we can see clear spaces between the character components. The line graphs drawn in Fig. 5.3(h) for the image in Fig. 5.3(g) shows that the enhancement step removes background noises clearly.

In this work, we prefer to use the above combination rather than popular canny edge operator because canny operator which is highly sensitive and outputs thinned edge images which are spurious in nature. Also we can see sobel and canny edge images appear with double lines. We can see that Sobel operator does not produce much spurious edges as in Fig. 5.3(i) compared to canny operator as in Fig. 5.3(k). This is because canny operators are very sensitive to Image pixels. The same thing can be confirmed from the line graphs in Fig. 5.3(j) and Fig. 5.3(l) where we can see more spikes in case of canny compare to sobel.Hence we decide that images are not smoothened using combination of sobel and canny as in Fig. 5.3(m). Due to this, intersection of these two operations will not yield good result in case of degraded documents such as Indus. Thus when compare to intersection of gradients as in Fig. 5.3(e) and Fig. 5.3(m) edges of gradient are very much sharpen compare to combination of sobel and canny. It can also be noticed in Fig. 5.3(f) some of the pulses are separated with certain wavelength whereas in Fig. 5.3(n) it is found more sharp spikes which are closed to each other. This signifies number of pixels in gradient image is reduced comparative to edge image. Other thing what we notice is, convolution of gradient image reduces discontinuities thus giving strong edges compare to sobel and canny edge combination. Thus, noisy pixels in gradients combination are also less. Double edges in combination of sobel and canny edge images decrease the spacing between the characters. Therefore, it can be concluded that the Sobel and the Laplacian combination is better for the Indus document.



(a) Gradient by Sobel operator

(b) Gradient by Laplacian operator

(c) Line graphs for the row shown in (a)

(d) Line graphs for the row shown in (b)

(e) Intersection of (a) and (b) gradient images

(f) Line graphs for the row in (e)

(g) Final cleaned image

(h) Line graphs for the row in (g)

(i) Sobel edge image for the image in Fig. 5.2(a)

(j) Line graphs for the row in (i)

(k) Canny edge image of the input image in Fig.5.2 (a)

(l) Line graphs for the row in

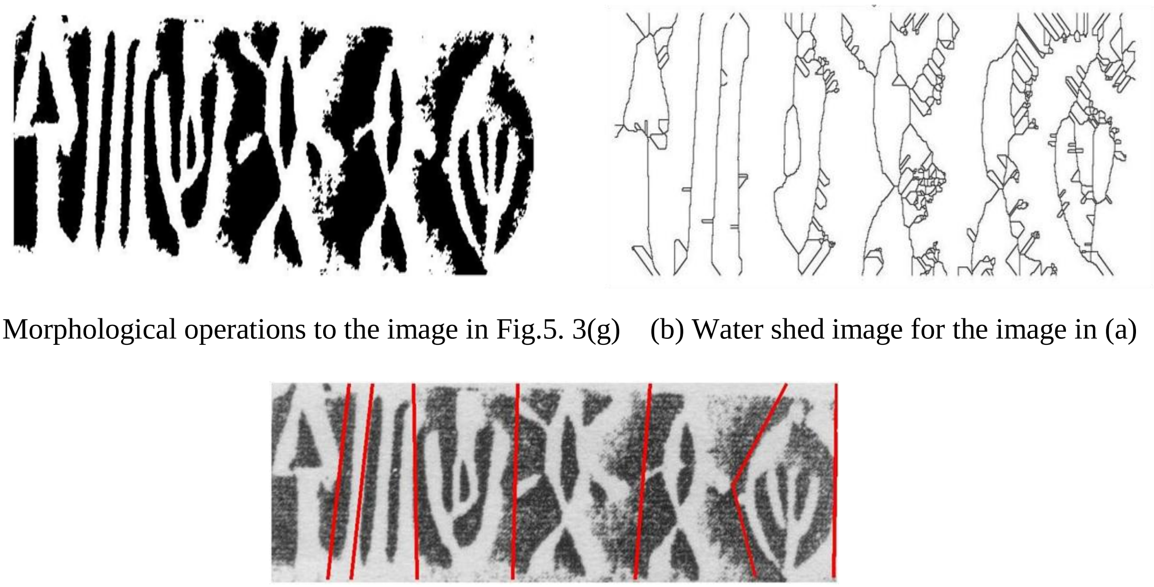
(m) Intersection of Sobel and Canny edge images

(n) Line graphs for the row in (m)

Fig.5. 3. Illustration of the proposed enhancement

5.2.2 Watershed Model for Character Segmentation

We know that water flows from maximum region to minimum region and stores in region called catchment basin. As observed in Fig. 5.3(g) there are many disconnected components in each of the text character. As water flows from maximal region to minimal region there are chances of water entering through these disconnections and forming the watersheds. Thus it is found necessary to group these disconnected components in each of the character. Thus we perform morphological operations as shown in Fig. 5.4(a) to group the disconnected components. It is noticed that edges of text pixels are cursive thus it is found pixels of single column not possible to segment characters. However water can be flowed at these cursive exterior edges of character. Hence this observation, inspired to use the watershed algorithm to find the boundaries of text characters. The watershed algorithm finds water flow and high volume of collection of water where there is a space between the character components. These two properties work well even if any touching exists between the character components. In this way, watershed algorithm helps in segmenting characters from Indus text line.



(a) Morphological operations to the image in Fig.5. 3(g) (b) Water shed image for the image in (a)

(c) Character segmentation

Fig. 5.4. Steps of the Proposed Character Segmentation

Watershed image for Fig. 5.4(a) is shown in Fig. 5.4(b).Watershed image shows some of the unclosed white regions where water can be stored. We know water flow from upward towards downward. It can be noticed there are small regions within each of the character component where small amount of water can be stored. There also exist regions between the text characters where large amount of water can be stored. These high amount storage regions clearly indicate spacing between characters. These region points are extracted and works well for segmentation of characters as shown in Fig. 5.4(c). Fixing bounding box for each of the connected components leads to overwhelming of boundaries as shown in Fig. 5.5. Hence the method finds the point where maximum water is stored. Further to detect the segmentation path we consider the region between the characters and count the number of white pixels between the extreme black pixels. We then compute the position of midpoint at the top, middle and bottom of the region called as segmentation path markers. Joining of these markers is the segmentation path as shown in Fig. 5.6.Thus input image of Fig. 5.2 (c) segmented using this proposed method is as shown in Fig. 5.4 (c).

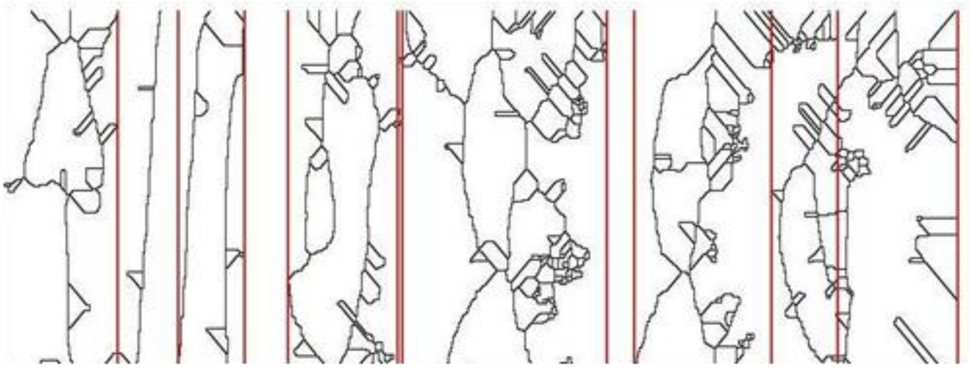


Fig. 5.5. Over segmentation

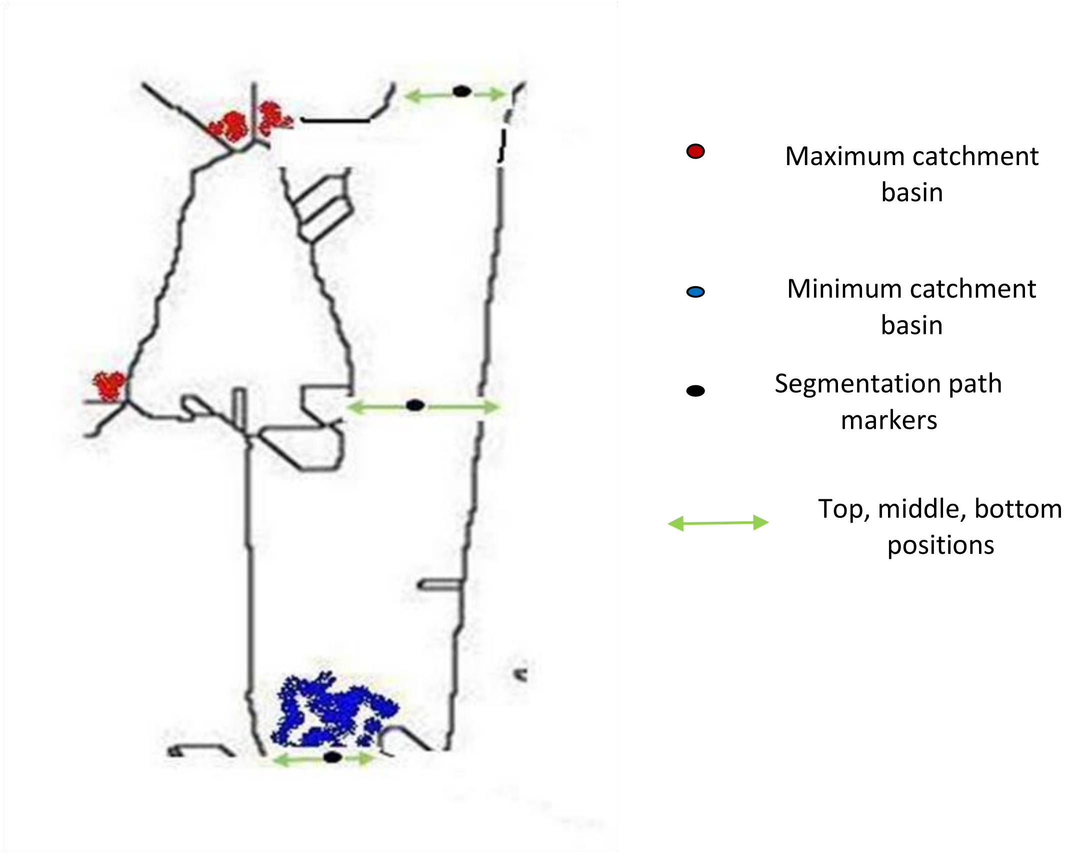
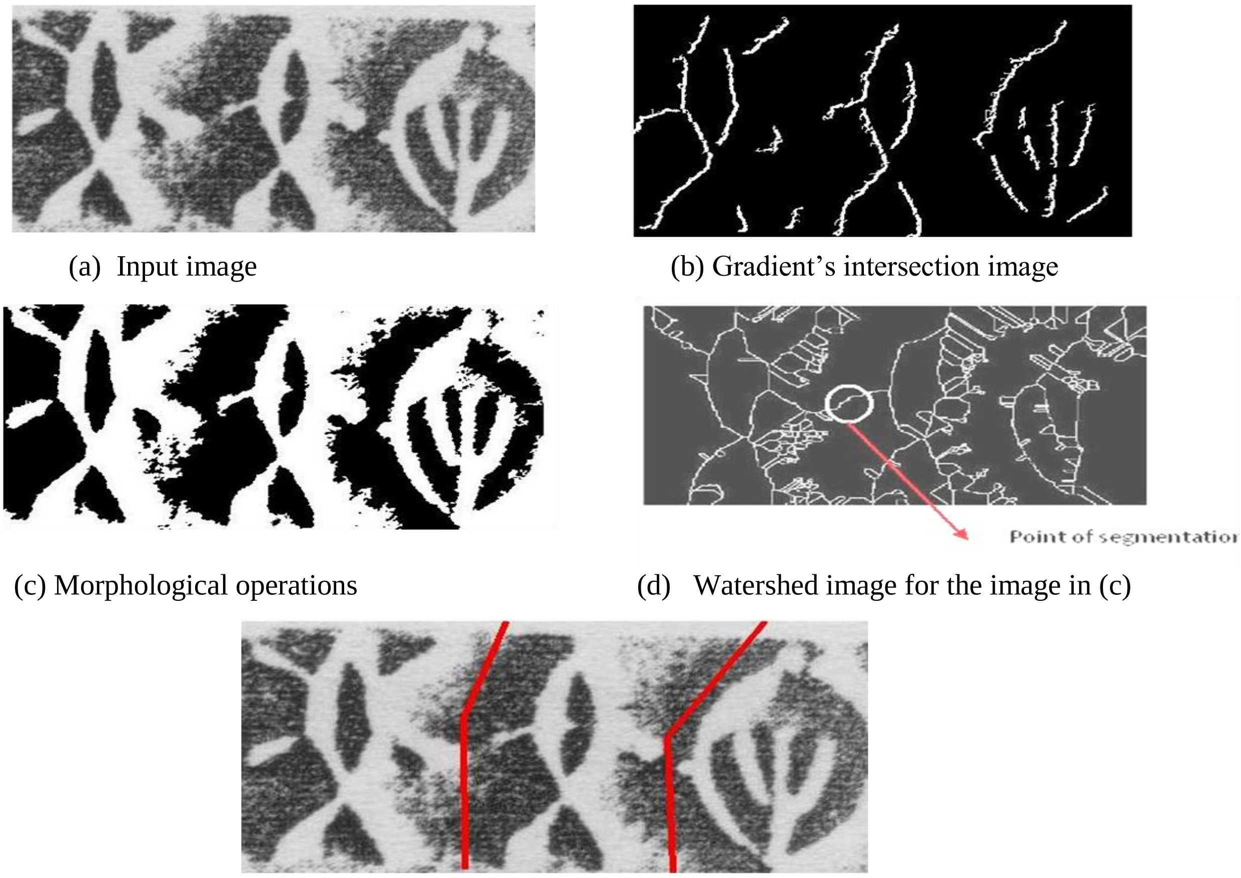


Fig. 5.6. Segmentation path detection between first 2 text characters of Fig. 5.4(b)

The effect of watershed algorithm can be seen in Fig. 5.4(c) that all characters are segmented correctly. One more illustration is presented in Fig. 5.7 where there is a touching exists between the characters. Fig. 5.7 shows that for the input image in Fig. 5.7(a), the proposed method obtain cleaned image by the previous step as shown in Fig. 5.7(b) and then the proposed method performs morphological operation to group the disconnected component as a single component as shown in Fig. 5.7(c). For the image in Fig. 5(c), when we apply watershed algorithm, there are chances of existing touching between character components due to complex background of Indus text line as shown in Fig. 5.7(d). In this case, when the proposed method estimates flow of water and volume of collection water, the space between the characters components has highest volume of the collection of water. Thus the markers detected guarantees the limits in which text characters exist. Therefore, the watershed algorithm segments characters components successfully for the case of touching characters as shown in Fig. 5.7(e).



(a) Input image

(b) Gradient's intersection image

(c) Morphological operations

(d) Watershed image for the image in (c)

(e) Character Segmentation

Fig. 5.7. Illustration for Proposed Character Segmentation for touching characters

5.3 Experimental Results

To evaluate the effectiveness of the proposed method we first conduct experiments on 100 text lines of English and 100 cursive scripts that includeTelugu, Tamil and Malayalam scripts. Experimental results for English text lines for the proposed and existing methods are shown in Table 5.1. Experimental results for other cursive scripts (Telugu, Tamil and Malayalam) for the proposed and existing methods are shown in Table 5.2.

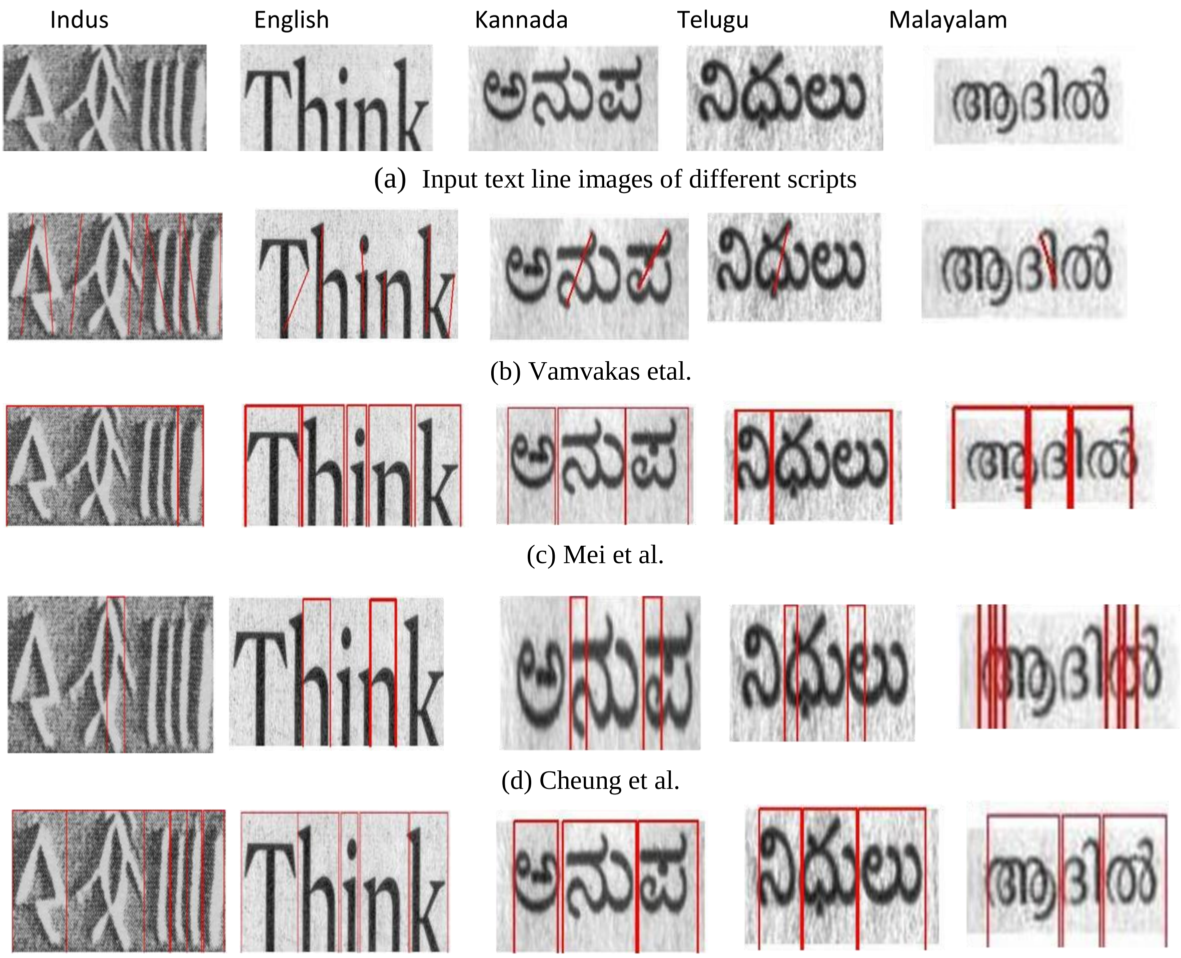
To conduct experiments on the proposed method for segmentation of characters we accumulate 500 Indus text line images segmented using method proposed in chapter 4.This is because scope of our research is limited to Indus scripts. This dataset includes a variety of text lines on different surfaces and different handwriting with different tools. As a result, this dataset is said to be complex compared. Fig. 5.4 and Fig. 5.7 show the results using the proposed method for untouching and touching characters. To measure the performance of the proposed character segmentation method, we use the well-known measures, namely, recall and precision and F-measure as defined in equation (5.6)-(5.8), where the test outcome may be in various facts such as TP (true positive) defined as the number of characters that are correctly segmented, Fp (false positive) is number of false characters detected as true, and Fn (false negative) is the number of characters which is true w.r.t false category. Experimental results for Indus text lines for the proposed and existing methods are shown in Table 5.3.

$$Recall = \frac{Tp}{(Tp + Fn)} \tag{5.6}$$

$$Precision = \frac{Tp}{(Tp + Fp)} \tag{5.7}$$

$$F\_Score = \frac{2Tp}{(2Tp + Fp + Fn)} \tag{5.8}$$

To show the effectiveness of the proposed method, we implement three existing methods for comparative studies. Sample qualitative results of the proposed and existing methods are shown in Fig. 5.8 for the different scripts text lines. Existing method Vamvakas et al.’s method does not give good result compare to propose method. Method finds segmentation points using feature points and the distances between feature points. The method gives poor results for Indus and other scripts such as English and south Indian scripts. Main reason for the poor result is that the method is not tolerant to degradation existed in Indus and scanned images of English and south Indian scripts. This is because scanned images have many disconnected pixels that are considered as true feature points in the method. Another existing method proposed by Me et al. uses connected components analysis and boundary of segmentation obtained using vertical projection profile for character segmentation. The method is tolerant to degradations to certain level. Hence compare to Vamvakas et al.’s method it gives good result for English script. Since the method finds the connected components for scripts. South Indian scripts and Indus scripts are more cursive hence there exists more number of connected components than the number of characters. This is due to background and structure of scripts. Similarly, Cheung et al.'s method does not segment characters properly because the proposed features extracted are based on the structure of a specific script. It segments using vertical projection profiles. This method gives the low values for the measures. However, there is some improvement in precision value of south Indian scripts in Table 5.2 comparatively. This is because Malayalam scripts do not have more sub components for the texts. Hence this gives more true positives comparatively. However, for English it gives low result since the text characters are almost have equal size. Thus experimental results shown for the proposed method in Table 5.1 to Table 5.4 gives good results for Indus and other script text line images because of the advantage of the enhancement step and the watershed model.



(a) Input text line images of different scripts

(b) Vamvakas etal.

(c) Mei et al.

(d) Cheung et al.

(e) Proposed method

Fig. 5.8. Qualitative results of the proposed method for Character segmentation

It is found necessary to evaluate the performance of the proposed method without enhancement to show the importance of enhancement. We evaluate the performance of the proposed method without enhancement using same quantitative measures, such as recall, precision and f-measure for different types of script data, we calculate the measures for English, South Indian scripts and Indus as reported in Table 5.1-Table 5.4, respectively. It is noted from Table 5.1-Table 5.4 that the proposed method achieves better results compared to the existing methods in terms of recall, precision and f-measure. When we compare the result of the proposed method without enhancement English scripts in Table 5.1 gives good result when compare to other scripts including Indus. This is because English scripts usually have less disconnected components. Hence there exists clear spacing between the characters. Watershed images for English text lines have maximum catchment basins in between the text characters. Therefore, compare to other scripts English scripts works well without enhancement.

Proposed method with enhancement gives good results for scripts like Kannada, Telugu, and Malayalam as shown in Table 5.2. Proposed method merges the small regions that are aroused due to presence of modifiers. This avoids the free region spacing between character component and modifier component. Moreover, in scanned images there appears the image of some false pixels. Thus it is needed to enhance the gradients of text pixels. Thus enhancement step in proposed method enhance text pixels and creates the segmentation regions between the characters. Thus with enhancement the proposed method gives good result not only for English it also gives good segmentation result for cursive scripts also. Reason is proposed watershed model is good for irregular shape scripts. Proposed method performance is low without enhancement. This is due to appearance of more noisy speckles gives more ridges in watershed model. Hence this increase number of catchment basin. Thus segmentation paths found are more than the number of characters. Therefore, to achieve good results for all types of scripts, we need the enhancement step to clean background before applying the watershed model. In this way, the proposed enhancement step contributes well for achieving good character segmentation results in this work.

Table 5.1. Performance of the proposed and existing methods on English script data

Methods	Recall	Precision	F_Measure
Mei et. al.	0.95	100	0.97
Vamvakas et. al.	0.3	0.375	0.33
Cheung et. al.	0.18	0.33	0.235
Proposed Method without Enhancement	0.84	0.875	0.85
Proposed Method with Enhancement	**100**	**100**	**100**

Table 5.2. Performance of the proposed and existing methods on other script (Kannada, Telugu, and Malayalam) data

Methods	Recall	Precision	F_Measure
Mei et. al.	0.3	0.5	0.375
Vamvakas et. al.	0.3	0.33	0.315
Cheung et. al.	0.12	0.54	0.196
Proposed Method without Enhancement	0.4	0.5	0.44
Proposed Method with Enhancement	**0.95**	**0.97**	**0.96**

Table 5.3. Performance of the proposed and existing methods on Indus script data

Methods	Recall	Precision	F_Measure
Mei et. al.	0.016	0.047	0.024
Vamvakas et. al.	0.016	0.024	0.019
Cheung et. al.	0.006	0.1	0.0116
Proposed Method without Enhancement	0.03	0.285	0.059
Proposed Method with Enhancement	**0.99**	**100**	**0.99**

Table 5.4. Overall performance of the proposed and existing methods on the whole data

Methods	Recall	Precision	F_Measure
Mei et. al.	0.37	0.92	0.5
Vamvakas et. al.	0.04	0.36	0.072
Cheung et. al.	0.2	0.32	0.147
Proposed Method without Enhancement	0.3	0.65	0.41
Proposed Method with Enhancement	**0.98**	**0.99**	**0.98**

5.4 Summary

In this chapter, we have proposed a new method for segmentation of characters from text lines in degraded historical document images like Indus. The proposed method explores the combination of Laplacian and Sobel operations for enhancing low contrast pixels in images by suppressing background noises. The characteristics of text components in an enhanced image are studied to eliminate unwanted background noise components, which results in a cleaned image with only edges which represent text components. We have proposed the watershed model for identifying spacing between characters by exploiting catchment basin and flow of water. Experimental results and the comparisons with the existing methods show that the proposed method outperforms the existing methods in terms of recall and precision. Our future plan would be extending for multiple touching character component images in multi scales or multi oriented environments.

Chapter 6

Character recognition

6.1 Background

In chapter 5, we have proposed methods for segmentation of Indus characters from text line images. Since characters from Indus document suffer from distortion, low contrast and complex background, recognizing Indus character is challenging. In this chapter, Inspired by Histogram Oriented Gradients (HOG) which is well known descriptor for object recognition, we propose a new Histogram Oriented Tangents (HOT) descriptor for Indus character recognition. The feature extracted by HOT are passed to SVM classifier or recognition.

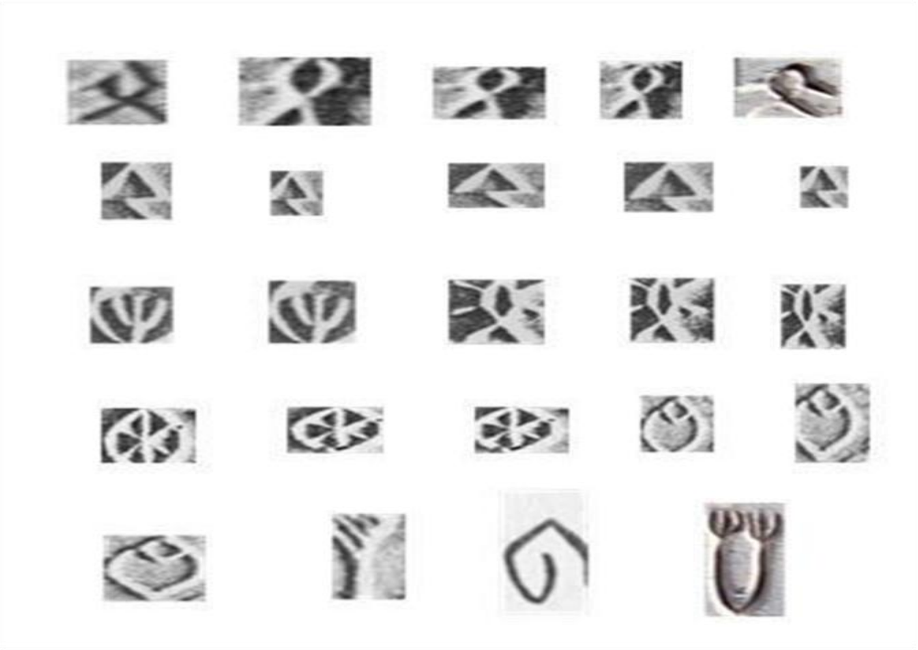


Fig. 6.1. Sample Indus character images

Some parts of the material of this chapter appeared in the following research papers:

1. "New Histogram Oriented Tangent Descriptor for Historical Indus Character Recognition System",(Under revision)

6.2 Proposed Methodology

Sample Indus character images are shown in Fig. 6.1 where we can see that character images are affected by multiple adverse factors such as low contrast, complex background, distortion, etc. Therefore, Indus character recognition is challenging and interesting. For a given Indus character image, the proposed method divides image into equal number of blocks. For each block, a new descriptor called Histogram Oriented Tangents (HOT) draws tangents for every pixel on boundaries. Further, we concatenate histograms of each blocks of the image to generate feature matrix. The proposed method use SVM classifier for feature matrix to recognize the characters. The steps of the proposed method are shown in Fig .6.2.

Thus the proposed method is divided into three subsections. Boundary fixing of the characters for block division is discussed in section 6.2.1. The method of finding the slope of each pixel using HOT descriptor is discussed in section 6.2.2. Slope based features based on run is discussed in section 6.2.3.

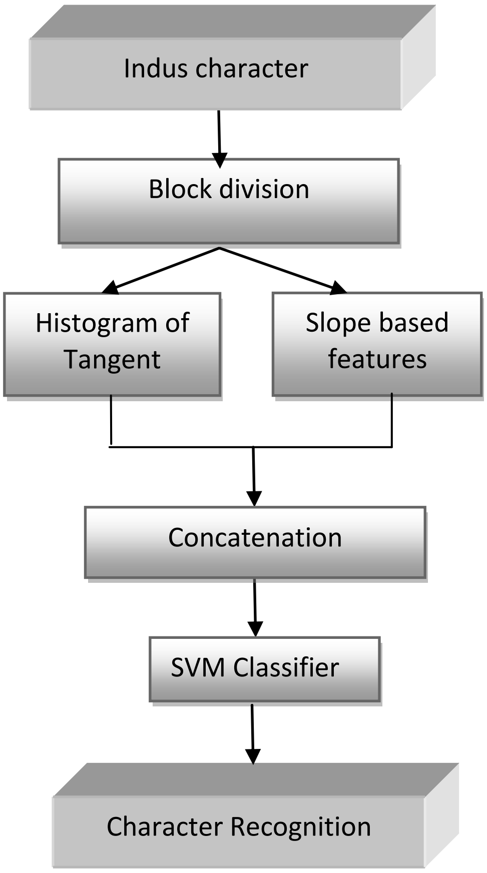


Fig. 6.2. Flow chart of the proposed method

6.2.1 Boundary Fixing for the Character Components

For each character image, the method obtains skeleton of the image. Then it divides the whole image into equal sized blocks by finding two farthest points vertically and horizontally. Two farthest points in vertical direction provide height of the character image and two farthest points in horizontal direction provides width of the character image as shown in Fig. 6.3(a) where we can red and yellow color pixels which represent two farthest vertical points and two farthest horizontal points, respectively.

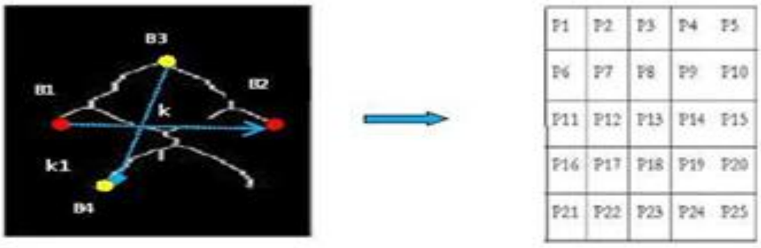


Fig. 6.3. Dividing character image into variable blocks

The formal algorithmic steps for finding width distance by scanning rows and scanning columns to find height is given below. Extreme lowest and highest points for row and columns are defined as in equation (6.1)-(6.4).

Let R be the rows of image $R = \{R_i\ , i = 1 \dots N\}$. $S_1$is the vector containing the position of first pixel encountered in each row and S_2 be the vector containing the position of last pixel encountered in each Column. B$_1$ and B$_2$ are the minimum and maximum column numbers.

$$S_1 = \{\mathrm{Col}_{1\mathrm{min}}, \mathrm{Col}_{2\mathrm{min}}, \dots, \mathrm{Col}_{\mathrm{nmin}}\} \text{ and } \mathrm{S}_2 = \\ = \{\mathrm{Col}_{1\mathrm{ma}}\mathrm{x}, \mathrm{Col}_{2\mathrm{max}}, \dots, \mathrm{Col}_{\mathrm{nmax}}\}$$

$$B1 = \sum_{i=1}^{N} Min(S_1)\ and\ B2 \\ = \sum_{i=1}^{N} Max(S_2) \qquad (6.1)$$

For each column of image, Let C={C$_1$,C$_2$,…,C$_N$} be the columns of image , S$_3$ be the vector containing the position of first pixel encountered in each column and S$_4$ be the vector containing the position of last pixel encountered in each column. B3 and B4 are the minimum and maximum row numbers.

$$S_3 = \{\mathrm{Row}_{1\mathrm{min}}, \mathrm{Row}_{2\mathrm{min}}, \dots, \mathrm{Row}_{\mathrm{nmin}}\}\text{and } \mathrm{S}_4 \\ = \{\mathrm{Row}_{1\mathrm{max}}, \mathrm{Row}_{2\mathrm{max}}, \dots, \mathrm{Row}_{\mathrm{nmax}}\}$$

$$B3 = \sum_{i=1}^{N} Min(S3)\ and\ B4 = \sum_{i=1}^{N} Max(S4) \qquad (6.2)$$

From Equations (6.2) and (6.3), boundary points for partitioning of pixels are calculated as in Equation (6.4) and (6.5).

$$Col_{width} = k = \frac{(B2 - B1)}{5} \qquad (6.3)$$

$$Row_{width} = k1 = \frac{(B4 - B3)}{5} \qquad (6.4)$$

Advantage of this procedure over conventional normalization is that the proposed block division preserves entire structure of character image. For each block Pi, Xi gives the width and Yi gives the height of each block. Formula using boundary pixels for computing Dimension of blocks [p1…p25] is determined using equations shown in Fig. 6.4.

Block number	Width	Height
P1	X1=$B3: (B3 + K1) - 1$	Y1=$B1: (B1 + K) - 1$
P2	X2=$B3: (B3 + K1) - 1$	Y2=$(B1 + K): (B1 + 2K) - 1$
P3	X3=$B3: (B3 + K1) - 1$	Y3=$(B1 + 2K): (B1 + 3K) - 1$
P4	X4=$B3: (B3 + K1) - 1$	Y4=$(B1 + 3K): (B1 + 4K) - 1$
P5	X5=$B3: (B3 + K1) - 1$	Y5=$(B1 + 4K): (B1 + 5K) - 1$
P6	X6=$(B3 + K1): (B3 + 2K1) - 1$	Y6=$B1: (B1 + K) - 1$
P7	X7=$(B3 + K1): (B3 + 2K1) - 1$	$Y7 = (B1 + K): (B1 + 2K) - 1$
P8	X8=$(B3 + K1): (B3 + 2K1) - 1$	Y8=$(B1 + 2K): (B1 + 3K) -$

		1
P9	X9=$(B3+K1):(B3+2K1)-1$	Y9=$(B1+3K):(B1+4K)-1$
P10	X10=$(B3+K1):(B3+2K1)-1$	Y10=$(B1+4K):(B1+5K)-1$
P11	X11=$(B3+2K1):(B3+3K1)-1$	Y11=$B1:(B1+K)-1$
P12	X12=$(B3+2K1):(B3+3K1)-1$	Y12=$(B1+K):(B1+2K)-1$
P13	X13=$(B3+2K1):(B3+3K1)-1$	Y13=$(B1+2K):(B1+3K)-1$
P14	X14=$(B3+2K1):(B3+3K1)-1$	Y14=$(B1+3K):(B1+4K)-1$
P15	X15=$(B3+2K1):(B3+3K1)-1$	Y15=$(B1+4K):(B1+5K)-1$
P16	X16=$(B3+3K1):(B3+4K1)-1$	Y16=$B1:(B1+K)-1$
P17	X17=$(B3+3K1):(B3+4K1)-1$	$Y17=(B1+K):(B1+2K)-1$
P18	X18=$(B3+3K1):(B3+4K1)-1$	$Y18=(B1+2K):(B1+3K)-1$
P19	X19=$(B3+3K1):(B3+4K1)-1$	$Y19=(B1+3K):(B1+4K)-1$
P20	X20=$(B3+3K1):(B3+4K1)-1$	$Y20=(B1+4K):(B1+5K)-1$
P21	X21=$B3+4K1):(B3+5K1)-1$	Y21=$B1:(B1+K)-1$
P22	X22=$(B3+4K1):(B3+5K1)-1$	$Y22=(B1+K):(B1+2K)-1$
P23	X23=$(B3+4K1):(B3+5K1)-1$	$Y23=(B1+2K):(B1+3K)-1$
P24	X24=$(B3+4K1):(B3+5K1)-1$	$Y24=(B1+3K):(B1+4K)-1$
P25	X25=$(B3+4K1):(B3+5K1)-1$	$Y25=(B1+4K):(B1+5K)-1$

Fig. 6.4. Width and height calculation of blocks

6.2.2 Histogram Oriented Tangent (HOT) Descriptor

Inspired by the work (Khare et al, 2015) where the descriptor is proposed for classifying text and non-text pixels, we introduce a new descriptor called Histogram Oriented Tangent (HOT) rather than moments.

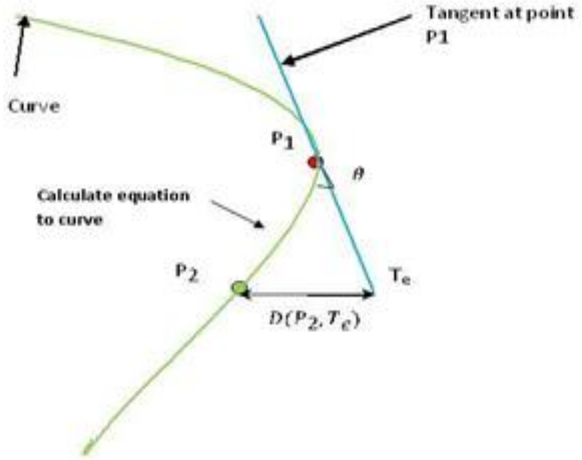


Fig. 6.6. Tangent angle estimation

For a pixel in each block, the proposed method calculate tangent angle as shown in Fig. 6.6. Then we study the pattern of angle information given by tangent, to extract shape of the character. As illustrated in Fig. 6.5, we can see P1 denotes a point for which the tangent is drawn and thus its angle is estimated. It is known that slope of tangent to a straight line is equal to line itself. However, an Indus character with irregular structure gives different tangent angles. After calculating angles for every pixel of the boundary of the character, we perform histogram operation for angular features of each block of the character image with four bins, namely, Horizontal, Vertical, the Secondary diagonal and the Principal diagonal. The proposed method concatenates the histograms of all the 25 blocks to obtain the final feature vector as shown in Fig. 6.6. In this way, the proposed HOT and invariant blocks division helps in extracting good features to recognize complex Indus characters.

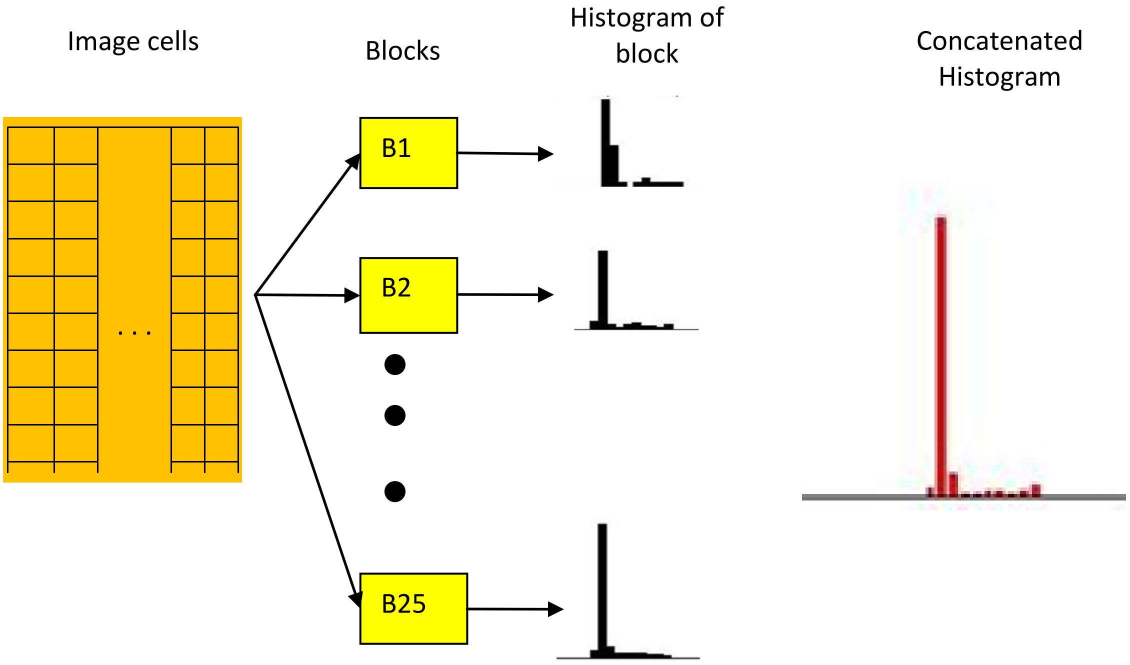


Fig. 6.6. Block diagram of HOT

Let $\{P1, P2, \dots, PN\} \in B$ be the boundary points (P) of block (B) of the character image. For eachP_i, its tangent is drawn and θ is estimated. How to find the tangent angle is the next question. Te is the point on tangent. For example to estimate θ at certain point $P1$ assume another point $P2$ at small distance apart from$P1$. Let us assume distance between $P1\ and\ p2\ is\ d1$; distance between $P1\ and\ Te\ is\ d2$; distance between $P2\ and\ Te\ is\ d3$. Therefore pixel point's $P1, P2, Te$ results in form of scalene triangle. θ is estimated using law of cosines as in equation 6.5.

$$\theta = Cos^{-1}\left(\frac{\mathrm{d1}^2 + \mathrm{d2}^2 - \mathrm{d3}}{2d1d2}\right) \tag{6.5}$$

Thus feature vector contains the angles of each block of the character image. We consider the following angle ranges for grouping Horizontal, Vertical, the Secondary diagonal and the Principal diagonal.

(i) Horizontal angle for pixels at 180 and 360. (ii) Vertical angle for pixels at 90 and 270 degrees, (iii) the Secondary diagonal angles for pixels at 315 and 225 degrees, and (iv) the Principal diagonal angle for pixels at 135 and 45 degrees. We plot a histogram for each block by encoding features as those four bins. Finally, we concatenate histograms of all the 25 blocks to obtain the final feature vector.

6.2.3 Slope based Features

To strengthen the HOT descriptor, we introduce new features which estimate slopes of lines in each block of the character image to study the structure of the character. For each block in the image, the proposed method identifies horizontal, vertical, the secondary diagonal and principal diagonal runs. The runs mean that the proposed method finds the continuous runs of pixels in the same direction. Various runs detected in the proposed method are horizontal, vertical, principal diagonal and secondary diagonal runs. Using these runs slope of the lines are estimated as shown in Fig. 6.7.Vertical runs of pixels comprise 90 and 270 degrees(Fig 6.8(a) and Fig 6.8(b)) ;Horizontal runs zero degree and 180 degree (Fig 6.8(c)); Principal diagonal 315 degrees (Fig 6.8(d)); Secondary diagonal 45 degrees (Fig 6.8(e)). In other words, the proposed method groups the whole image into horizontal, vertical, the secondary diagonal and the principal diagonal lines by using respective runs. As described in Section 6.2.2, we plot a histogram for the angles of each block with the above mentioned four bins. Then we concatenate bins of all the histograms of all the 25 blocks. Finally, the features are encoded as 0 for horizontal, 1 for vertical, 2 for the secondary diagonal and 3 for the principal diagonal to obtain a feature vector.

Horizontal line (H1, H2, H3, H4) Vertical line (V1, V2)

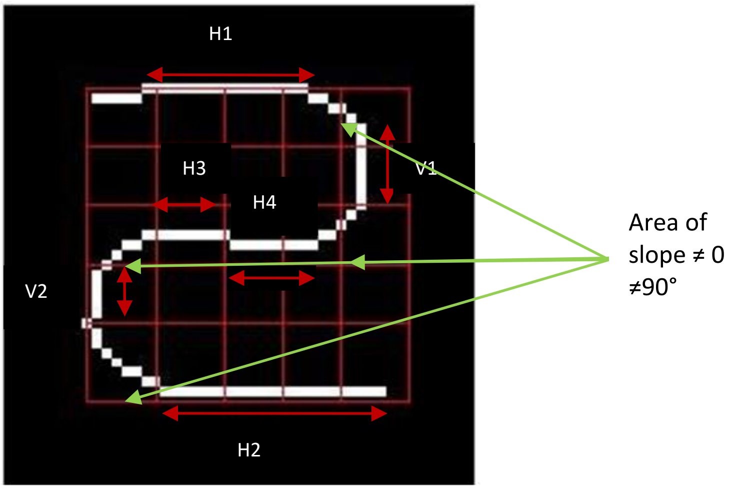


Fig. 6.7. Encoding angle information

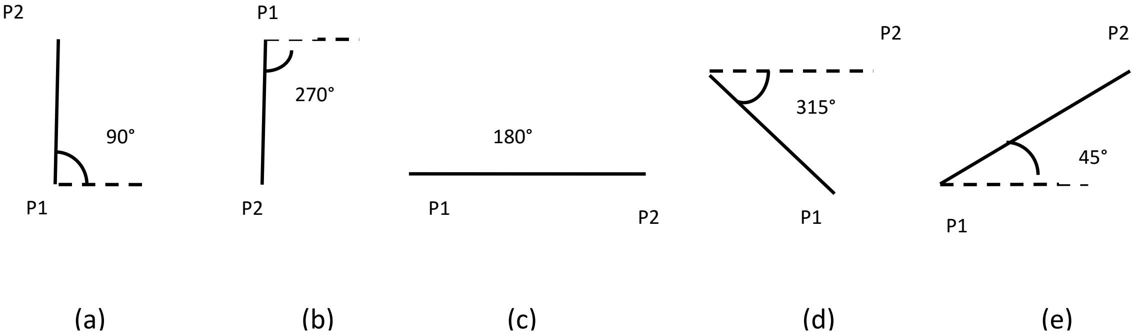


Fig. 6.8.Angle estimating using different slopes

For the feature vector containing the encoded information of the input image, the proposed method uses a Support Vector Machine (SVM) classifier for recognition. For experimentation, we use a 10 fold cross validation procedure to choose training and testing samples rather than choosing them manually. Since SVM with RBF kernel is capable to classify multi-classes, we intend to use the same SVM classifier for Indus character recognition in this work.

6.3 Experimental Results

To evaluate the proposed method, we use standard databases, namely, ICDAR 2011 (Kartzas et al, 2011) which includes scene characters with lots of font, background, font size variations, Street View Text (SVT) which is slightly complex compared to ICDAR 2011 as it involves characters from building, greenery, tree and sky backgrounds (Nguyen et al, 2014), and Char 74 k (De-Campos et al, 2009) which includes characters of multi-lingual with complex backgrounds. Apart from standard databases, we create an Indus character database collected from archeology survey of India and Mysore, which have lots variations in writing style, low contrast, poor visibility and complex background. This is much more complex than above standard databases. In summary, 1000 characters from ICDAR 2011, 1000 from SVT, 7000 from Ch 74k and 1000 from Indus databases are considered, which include 10000 characters for experimentation to measure the performance of the method. We consider standard character recognition rate for performance evaluations.

To show the proposed method is effective, we implement a few existing methods for comparative studies. The method proposed by Matsuda et al, 2014 is developed for character recognition from camera captured images. This method is capable of handling complex and plain background images. The method proposed by Kale et al, 2014 is developed for handwritten Devanagari compound character images. It has the ability to handle variations of different handwriting styles. The method proposed by Li et al, 2014 is developed for recognizing characters from historical images, where we can see degradations and distortion backgrounds. Apart from this, since the proposed method is similar to HOG method, we give comparative studies with HOG to show that the proposed method is better than HOG. We choose these four methods from different categories for comparative study because Indus characters may appear similarly to scene, handwriting, and multi-lingual. Therefore, for fair comparative studies, we choose the sate of art methods from each category.

The quantitative results of the proposed and existing methods are reported in Table 6.1, where one can notice that the proposed method outperforms the four existing methods in terms of character recognition rate. HOG gives the best results for Ch 74k database compared to the other existing methods because HOG is good for high contrast images, and Ch 74 data provides good contrast compared to ICDAR 2011, SVT and Indus data. Hence, it gives the best for Ch 74 k data. Kale et al. is the best for ICDAR 2011 compared to other existing methods because scene characters share the features of compound characters of Devanagari scripts. Li et al. score the best character recognition rate for Ch 74 k data compared to the other existing methods as Ch 74 k data involve multi-lingual data with high contrast background. Matsuda et al. achieved the best character recognition rate for Indus compared to the other existing methods because Indus characters share the similar complexity of scene characters. On the other hand, the proposed method is the best for all the four different databases. It is observed from the results of the existing methods that none of the method reports good results for all the four databases. The main reason is that the methods are developed for particular databases but not for multiple databases. Therefore, we can conclude that the proposed method which involves HOT, angular features from slopes and SVM classifier is capable of tackling the problems of Indus and other database as well.

Table 6. 1. Character recognition rate of the proposed and existing methods (in %)

Sl.no	Dataset	Number of samples	Existing Methods				Proposed Method
			Khare et al, 2015 (HOG)	Kale et al, 2014	Li et al, 2014	Matsuda et al, 2014	HOT
1	ICDAR 2011	1000	40.0	80.0	80.0	20.0	83.7
2	SVT	1000	20.0	40.0	70.0	35.0	80.0
3	Char 74k	7000	50.0	70.0	82.0	40.0	85.0
4	Indus	1000	20.0	20.0	20.0	80.0	97.0

Table 6.2 shows the performance analysis of the proposed and existing methods by varying the number of training samples. It can be seen when the methods are trained with variable samples proposed method gives good recognition rate when compare to other existing methods. Method HOG gives the good result for plain text images but not for our dataset. The reason is the variable gradient values in different images of similar texts. This is due to variation in inscribing Indus characters using hand tools. It can be seen from Table 6.2 processing time is also less compare to existing methods. It can be seen in Fig. 6.9 salt and pepper noise with threshold of 0.3 does not have much effect on edge pixels. Experimental results for existing and proposed a method

by varying noise at different thresholds is given in Fig.6.10.Whereas the proposed method is resistive to certain blurness. Experimental results shows that addition of blurriness as in Fig. 6.11 will not have much effect on edge pixels to compute HOT and slope based features. Fig 6.12 shows the performance of proposed and existing methods. Graph shows the good result when the blurness is in the range of 20…40. This is because we have thick edges in Indus compare to plain text images. When the blurness increases some pixels will be more effected there exists still, some of the white pixels at the edges that contributes for the proposed method. This shows the proposed method is tolerant to certain degradation.

Table 6.2. Character recognition rate (CR) and Processing time (PT) of the existing and proposed methods by varying the number of training samples (in sec)

Method	No of samples=50		No of samples=100		No of samples=300		No of samples=500	
	CR	PT	CR	PT	CR	PT	CR	PT
Khare et al, 2015 (HOG)	5.0	1.7	25.0	1.7	30.0	2.6	60.0	2.7
Kale et al, 2014	5.0	16.3	20.0	28.3	20.0	44.0	30.0	55.5
Li et al, 2014	4.0	74.2	10.0	181.2	15.0	504.2	20.0	573.9
Matsuda et al, 2014	20.0	61.3	40.0	171.1	60.0	493.7	75.0	503.7
Proposed Method (HOT)	67.0	0.4	78.0	0.6	85.0	2.4	98.0	2.5

Well known method of recognition using Histogram of gradients gives the magnitude and direction of gradients of pixels. Whereas Indus carved on hard materials have greater influence on gradients. This is because of uneven surface of background. Therefore along with the edge pixels the effect of background gradients will also be included. Gradients have effect for small changes. So, when the blurness increases intensity of pixels changes. This will have the greater effect on gradients.

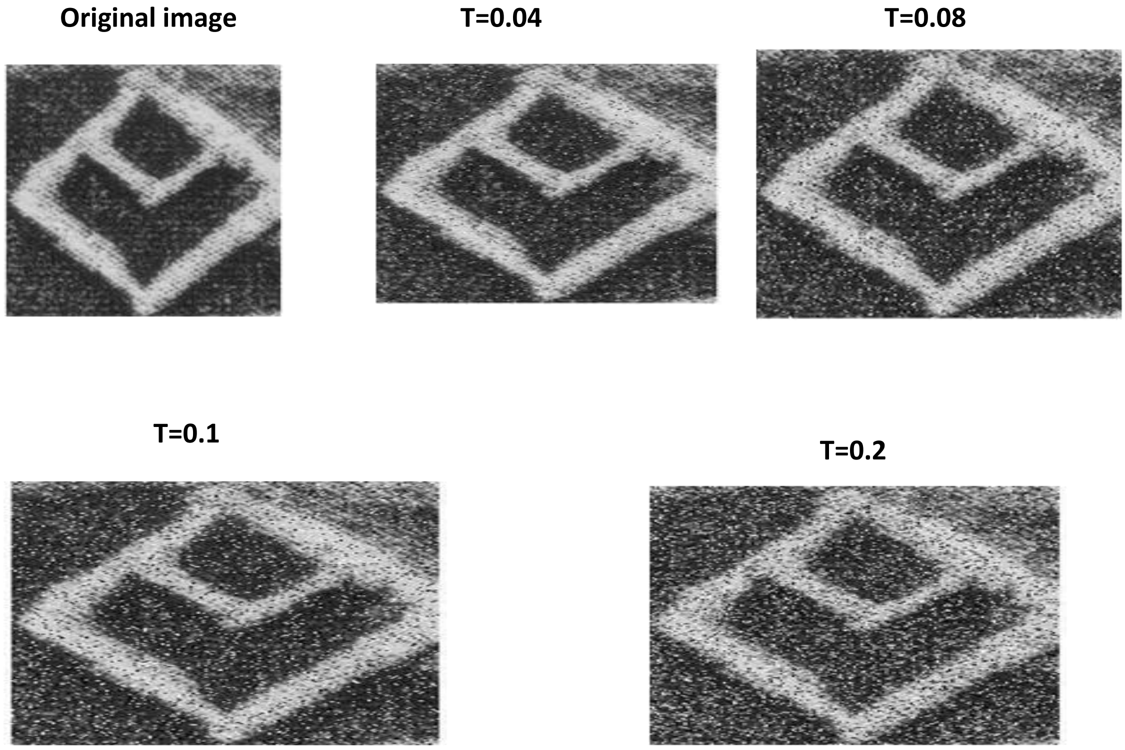


Fig. 6.9. Quantitative results for salt and pepper noise induced images with different thresholds (T) using proposed method

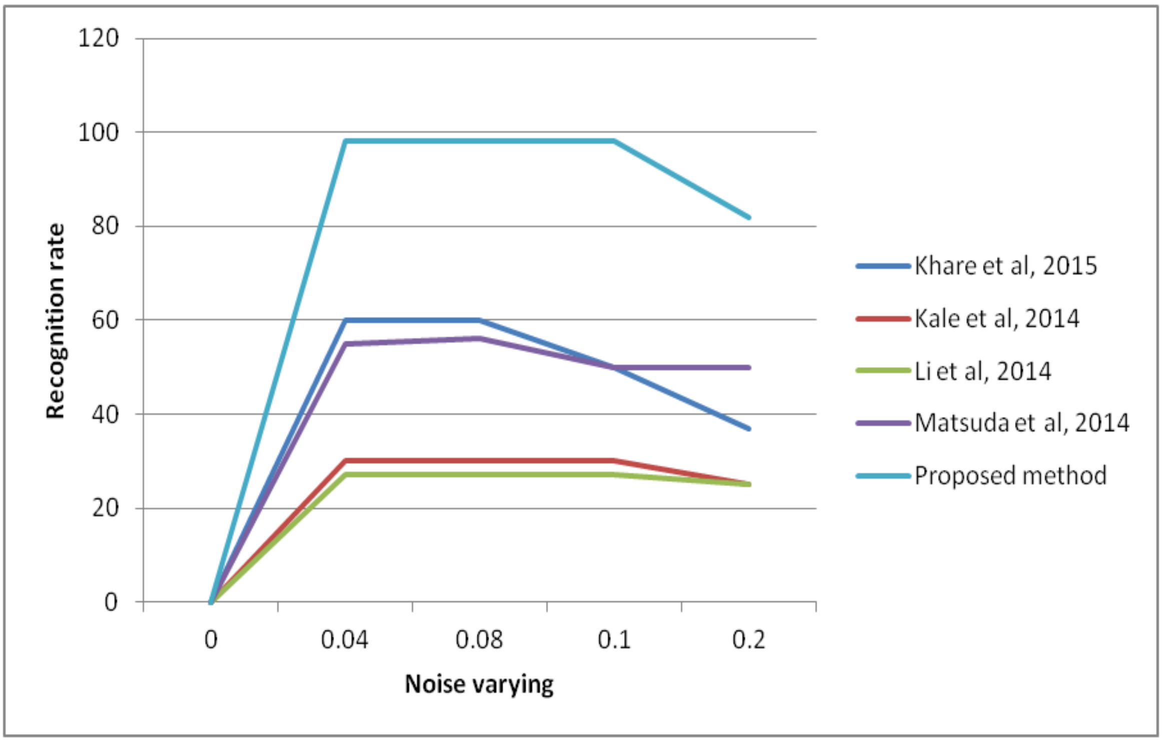
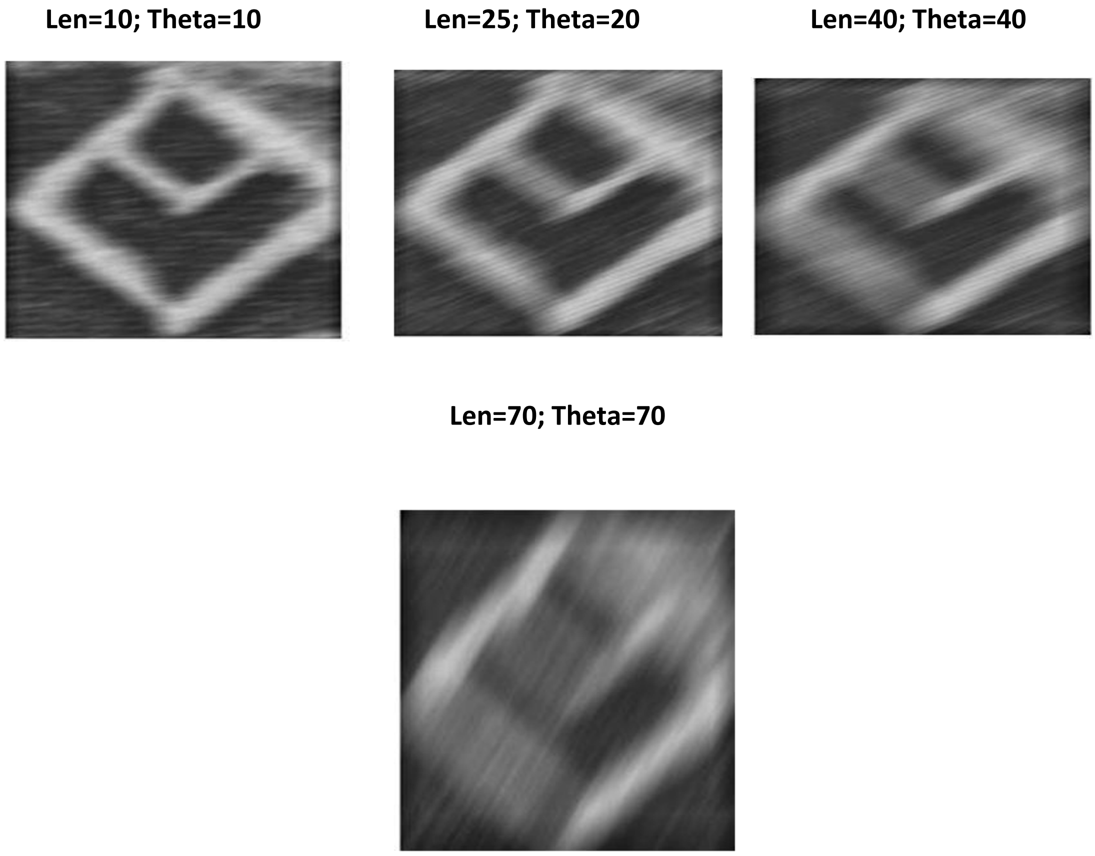


Fig. 6.10. Line graph showing recognition rate by varying Salt and Pepper noise for existing and proposed methods.



Len=10; Theta=10

Len=25; Theta=20

Len=40; Theta=40

Len=70; Theta=70

Fig. 6.11. Feature values of image with variable motion blurness

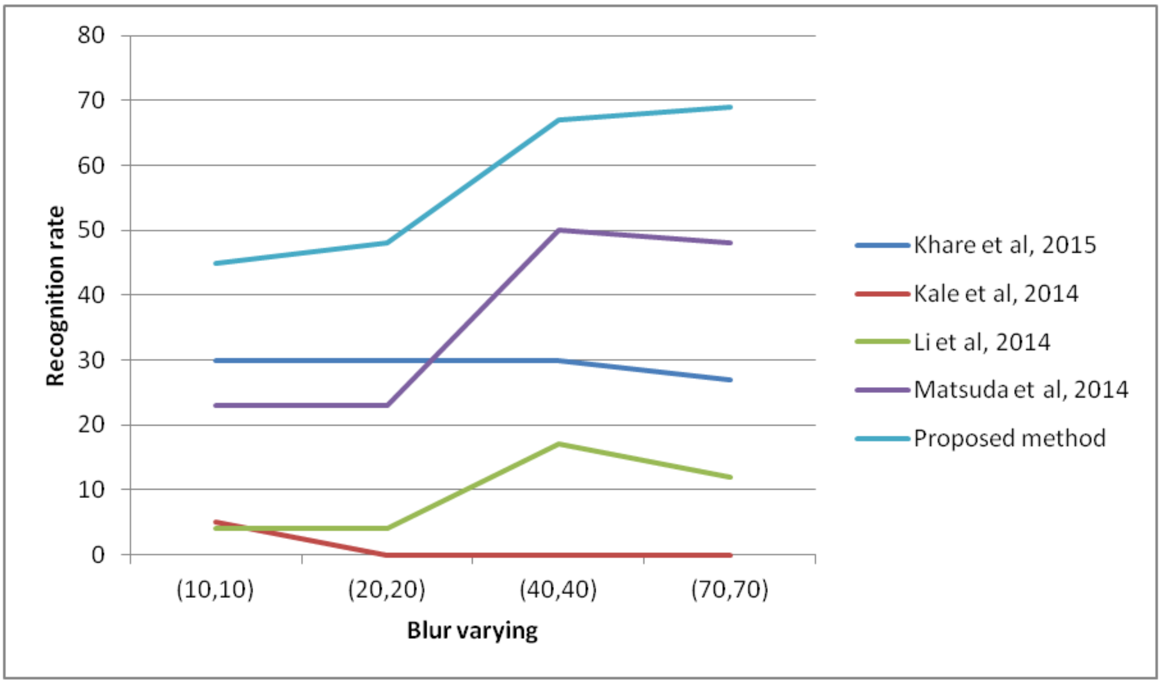


Fig. 6.12. Line graph showing recognition rate by varying motion blurness for existing and proposed methods.

6.4 Summary

We have proposed a new descriptor called Histogram Oriented Tangents (HOT) and angular features from slopes of character images for recognizing Indus characters in this chapter. Unlike HOG that considers orientations given by gradients, the proposed method considers orientations given by tangents at boundary pixels. The proposed method introduces a new idea for finding character regions which are used for dividing an input image into an equal number of blocks. In addition, the proposed method extracts angular features for finding slopes of strokes of character images to strengthen the discriminative power. We also perform histogram operation for the angular features given by HOT and slopes to quantize them into four bins, namely, Horizontal, Vertical, Secondary and Principal diagonals. The extracted features are encoded for the final recognition with the help of an SVM classifier. Experimental results on different databases and comparative studies with the existing methods show that the proposed method is better than the existing methods in terms of recognition rate. Our next target is to extend this method for recognizing characters in video and multi-scripts.

Chapter 7

Conclusion and future work

7.1 Summary

In this thesis, we proposed novel methods for historical Indus document understanding. This chapter presents summary of each chapter followed by limitations.

Chapter 1 provides introduction to Indus document and how Indus document is different from other documents. In addition, the importance of Indus document understanding especially for Archeological department point of view. It also discusses motivation, challenges, applications, major contribution of the proposed work.

Chapter 2 presents comprehensive literature survey on document understanding, script identification, text line segmentation, character segmentation and character recognition. Based on literature survey, we list out the challenges and unsolved issues.

Chapter 3 proposed different methods for identification of Indus scripts. We propose a method to identify Indus and English documents. Method is based on the study of straightness and cursiveness of the components. We use morphological operation to group the disconnected components as one to identify Indus and region growing for merging neighbor components to identify English documents. However this method gives more false positives thus we propose method based on skewness property for extracting straight and cursiveness of the components.

These methods are limited to identify scripts as three classes. Hence we propose a new method based on spatial relationship between dominant pixels to study structure of the components. This method is better than existing methods in terms of identification rate. Experimental results show that the proposed method works well for seven classes.

In order to identify the Indus document from other documents, Chapter 3 proposes methods based on statistical, structural, spatial and angular features. The scopes, motivation, limitations of the each method are discussed. In this chapter, the proposed method identifies Indus document from English, South Indian documents and few other Indian scripts.

Chapter 4 introduces nearest neighbor criterion for text line segmentation from the Indus document. Since the spacing between the characters is uniform, the proposed method explores this observation with the help of boundary fixing and boundary expanding in different direction.

Chapter 5 presents watershed model for character segmentation from text lines. The special features of watershed models, namely, water flow and catchment basin are used for finding spacing between the characters despite touching exist between the characters.

Chapter 6 proposes a new descriptor called Histogram Oriented Tangents (HOT) as inspired by Histogram Oriented Gradients (HOG) for feature extraction. The proposed method finds angle using tangent information and then form histogram for each block of the character image. The result of histogram concatenation of the give feature matrix. The feature matrix is passed to SVM classifier for recognition.

7.2 Future Work

Our literature study shows that methods on Indus document are still an open research problem due to its complexity. Hence we restricted our attention for developing novel methods for Indus document understanding. In depth literature study reveals that there are number of areas that are still left UN touched in case of Indus scripts. Hence we outline some of the points that we can consider in our future work.

Methods proposed in Chapter 3 are novel approaches for identification of Indus and other scripts. Methods proposed in this chapter are for three classes and seven classes. Our future work would be expanding the proposed method for other Indian and Foreign scripts.

A mathematical operation on gradient images for enhancement of Indus text and non text proposed in Chapter 4 is not 100 % free from unwanted components and noise. This is due to special structure of background and text. Hence further we extend work for preprocessing of Indus and other historical document images.

Method proposed in chapter 5 works well for non linear spacing of characters. The method also shows well performance for touching characters at single point. Thus our future plan would be extending the same method for blur images and multiple touching character component images in multi scales or multi oriented environments.

In chapter 6 we touched only slope based features. Since Indus characters are very complex still there are some of the hidden features that have to be exploited. Hence further we improve the performance of system by building visualization tools with these features. We also extend this method for recognizing characters in video and multi-scripts.

Due the irregular structure of background and characters in Indus image building Braille characters for these scripts is challenging. Hence we extend our work in building Braille characters for Indus scripts which helps to know the features of past characters for blind people.

Thus, we declare there are lots of untouched problems in this area and further there is scope in developing methods for future work to improve the accuracy and reducing the complexities of Indus scripts. Thus the proposed and future works helps for decipherment of Indus scripts which is still an open problem in the history of mankind.

List of Publications

International Journal Publications

1. A.S.Kavitha, P. Shivakumara, and G. Hemantha Kumar, "An Integrated Method for Classification of Indus and English Document Images",Lecture Notes in Electrical Engineering (LNEE), January 2014, Vol.248, pp.343-355.

2. A.S.Kavitha, P. Shivakumara, G. Hemantha Kumar and C. L. Tan, "A Robust Script Identification System for Historical Indian Document Images", Malaysian Journal of Computer Science(MJCS) , Vol. 28(4), 2015, pp 283-300 283.

3. A.S.Kavitha, P. Shivakumara, G. Hemantha Kumar and Tong Lu , "Text Line Segmentation in Degraded Historical Documents",Egyptian Informatics Journal,Vol 17(2),July 2016, pp 189-197.

4. A.S.Kavitha, P. Shivakumara, G. Hemantha Kumar and Tong Lu , "A New Watershed Model based System for Character Segmentation in Degraded Text Lines", International Journal of Electronics and Communications(AEU),Vol 71,2017, 45-52.

5. A.S.Kavitha, P. Shivakumara, G. Hemantha Kumar , Tong Lu and Chee Seng Chan, "New Histogram Oriented Tangent Descriptor for Historical Indus Character Recognition System", (Under revision)

Conference Publications

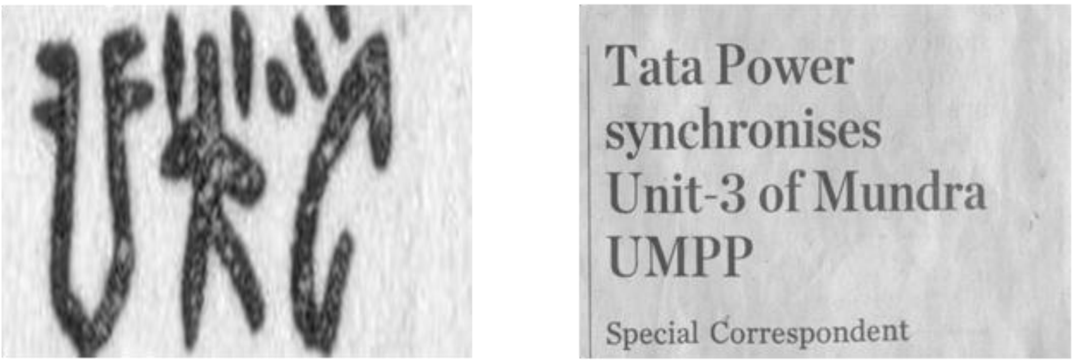
1. A.S.Kavitha, P. Shivakumara, and G. Hemantha Kumar, "An Integrated Method for Classification of Indus and English Document Images", International Conference on Emerging Trends in Electronics Computer Science and Technology (ICERECT-12), 21st December 2012.

2. A.S.Kavitha, P. Shivakumara, and G. Hemantha Kumar, “Skewness and Nearest Neighbour based Approach for Historical Document Classification”, Proceedings of the 2013 International Conference on Communication Systems and Network Technologies(CSNT, IEEE Computer Society Washington, DC, USA), April 2013, pp. 602-606.

Appendix

Though the methods proposed in earlier chapter gives effective results, but we fairly noticed some of the cases in which the proposed method fails to give appropriate result. Hence, in this section we report the different failure cases. These cases presented are the cases identified during experimentation of the proposed algorithm discussed in earlier chapters.

Different failure cases for the method proposed for identification of scripts by computing variance features for dominant pixels are shown in Fig A.1. However the method fails to detect the Indus document image in case when the image contain only text line as in Fig A.1(a).The method also fails to match the variance in case of documents with different size texts as shown in Fig A.1(b).



Tata Power synchronises Unit-3 of Mundra UMPP

Special Correspondent

(a) Indus text image identified as Hindi document

(b) English text image identified as Tamil document

Fig.A.1. Failure cases in identification of scripts

Proposed method for Text line segmentation in chapter 4 segments text images. Text line segmented also consists of non text as shown in Fig.A.2. Pruning is not effective in case of multiple contacts between text and non text as shown in Fig.A.2.Method do not give the exact boundary pixels for segmentation of text lines in case of degradation factor like edge brokings as seen in Fig.A.2.

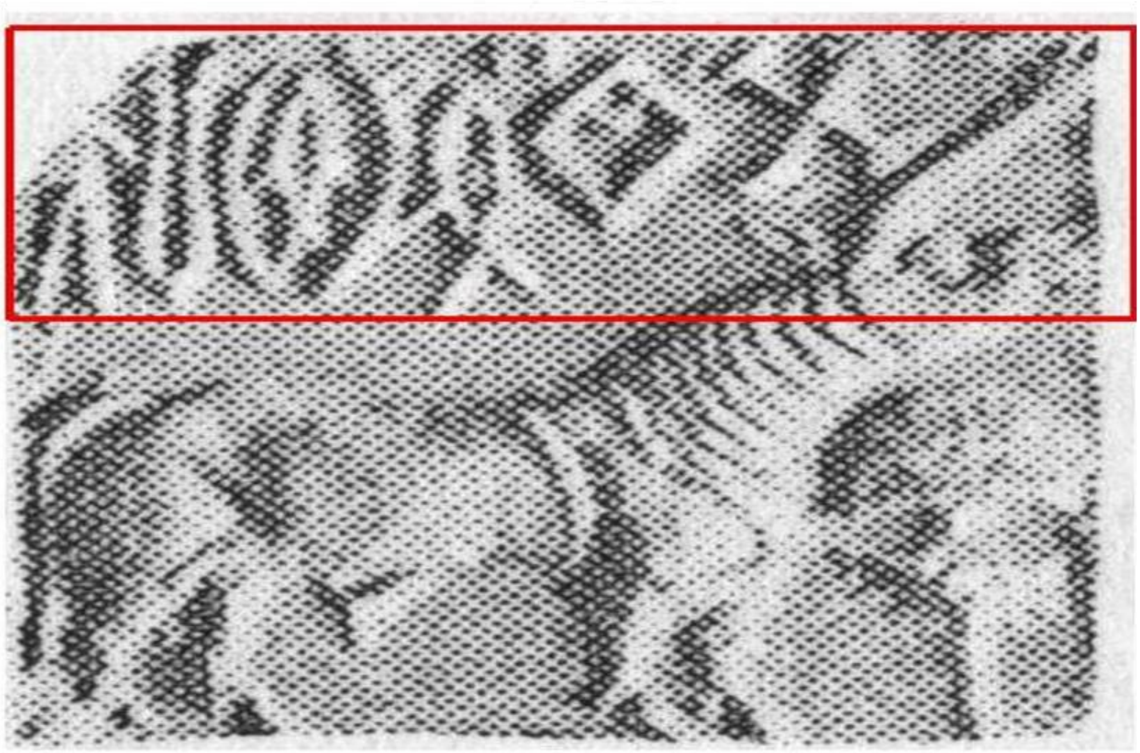
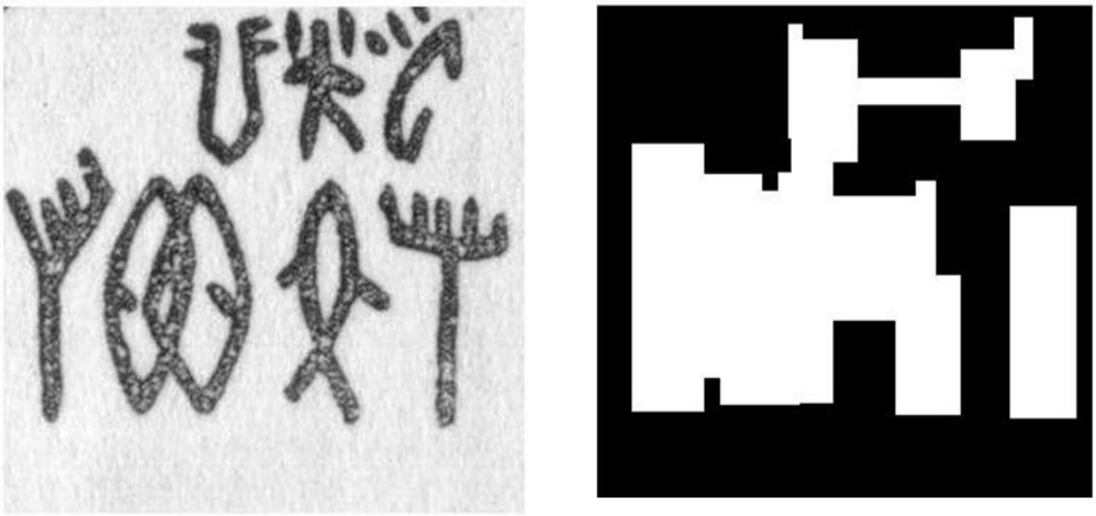


Fig. A.2. Over segmentation of text line in degraded Indus document image

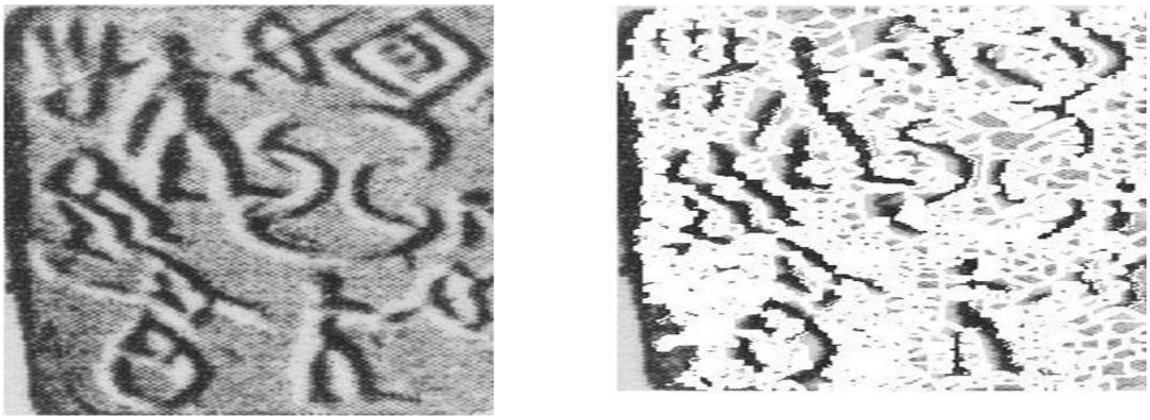
Text line segmentation method fails to form the right clusters in case of multiple text lines as shown in Fig.A.3. This is because spacing between the lines is lesser than space between the characters. Thus clusters formed by merging do not segment lines correctly.



(a) Input image (b) Clusters formed

Fig.A. 3. Failure of text line segmentation in case of multiple text lines in Indus images

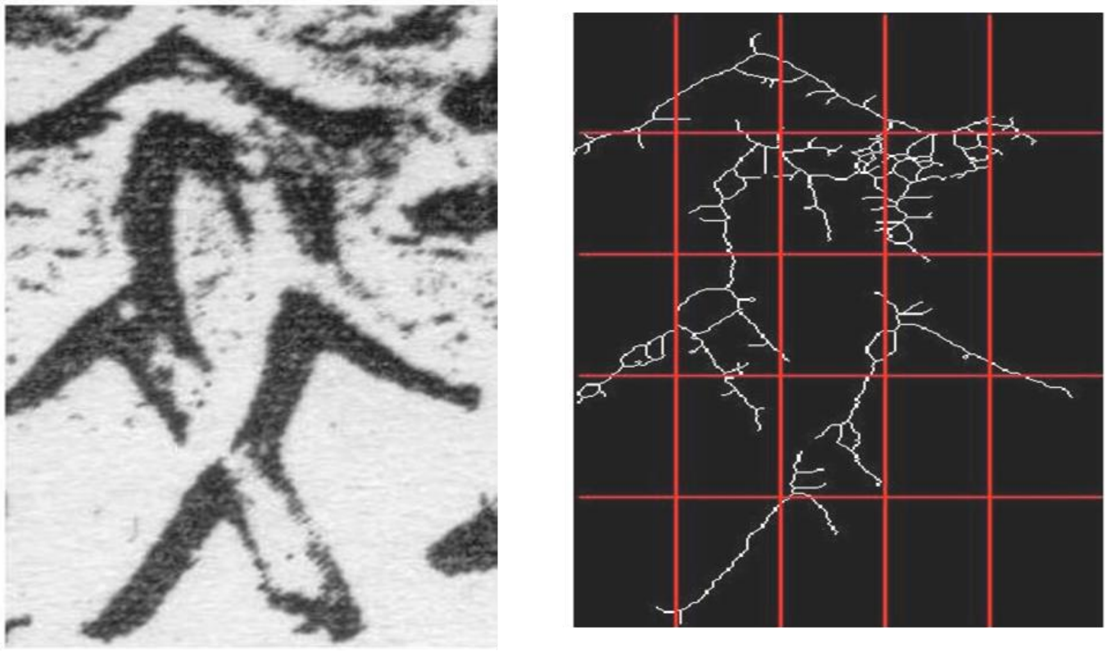
Proposed method for character Segmentation fails in case of when characters in text line are in high disorder as shown in Fig A.4 (a). The method dissolves the watershed lines as shown in Fig 4(b) with unclear spacing for segmentation.



(a) Indus image (b) Watershed image

Fig A.4. Failure of character segmentation in case of high disordering of texts

Proposed method for character recognition fails to determine run type in case of block containing loop edge and edge with multiple branches as shown in Fig.A.5 (b)



(a) Indus image (b) Image with block division

Fig.A. 5. Character recognition failure

Thus in this framework we listed the inverse cases of the proposed methods. Thus we conclude the proposed methods discussed conduct well in developing novel methods for Indus document understanding.

Bibliography

1. D. Ghosh , T. Dube and A.P. Shivaprasad , “ Script Recognition-Review”. *IEEE Transactions on Pattern Analysis & Machine Intelligence* ,Vol.32, December 2010 , pp.2142-2161.
2. T.N.Tan (1998). Rotation Invariant Texture Features and Their Use in Automatic Script Identification. *IEEE Transactions on* Pattern Analysis and Machine Intelligence.Vol. 20, No.7, *July 1998* pp.751-756.
3. A. Busch, W. W. Boles and S. Sridharan, Texture for Script Identification, IEEE Transactions on PAMI, 2005, pp 1720-1732.
4. L. Shijian and C.L. Tan, “ Script and Language Identification in Noisy and Degraded Document Images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,Vol.30, No. 1, January 2008, pp. 14-24.
5. S. Jaeger, H. Ma and D. Doermann, Identifying Script on Word-Level with Informational Confedence, In Proc. ICDAR, 2005, pp 416-420
6. P. B.Pati, and A. G. Ramakrishnan , “Word level multi-script identification”. *Pattern Recognition Letters*, Elsevier, Vol. 29, No. 9, July 2008, pp.1218-1229.
7. S. Chanda, S. Pal, K. Franke and U. Pal, Two-stage Approach for Word-wise Script Identification, In Proc. ICDAR, 2009, pp 926-930.
8. S. Chanda, O. R. Terrades and U. Pal, SVM Based Scheme for Thai and English Script Identification, In Proc. ICDAR, 2007,pp 551-555
9. L. Li, and C. L. Tan, “Script Identification of Camera-based Images”. *International Conference on Pattern Recognition*, IEEE, December 2008, pp. 1 – 4.
10. A. M. Namboodiri and A. K. Jain, On-line Script Recognition, In Proc. ICPR, 2002, pp 736-739.
11. U. Pal. S. Sinha and B. B. Chaudhuri, Multi-Script line identification from Indian documents, In Proc. ICDAR, 2003, pp 880-884.

12. S. Choudhury, G. Harit, S. Madnani and R. B. Shet, Identification of Scripts of Indian Languages by Combining Trainable Classifiers, In Proc. ICVGIP, 2000.
13. S. Chanda and U. Pal, English Devanagari and Urdu Text Identification, In Proc. ICDAR, 2005, pp 538-545.
14. L. Zhou, Y. Lu and C. L. Tan, Bangla/English Script Identification based on Analysis of Connected Component Pro_les, In Proc. DAS, 2006, pp 243-254.
15. R. Gopakumar, N. V. Subbareaddy, K. Makkithaya and U. D.Acharya, Zone-based Structural features extraction for Script Identification from Indian Documents, In Proc. ICIIS, 2010, pp 420-425.
16. P. K. Aithal, G. Rajesh, D. U. Acharya, M. Krishanamoorthi and N. V. Subbareddy Script Identification for a Tri-lingual Document, In Proc. CNC, 2011, pp 434-439.
17. R. Rani, R. Dhir and G. S. Lehal, Comparative Analysis of Gabor and Descriminating Features Extraction Techniques for Script Identification, In Proc. ICISILL, 2011, pp 174-179.
18. S. Ghosh and B. B. Chaudhuri, "Composite Script Identification and Orientation Detection for Indian Text Images", In Proc. ICDAR, 2011, pp 294-298.
19. P. Shivakumara, T. Q. Phan, Z. Ding, S. Lu and C. L. Tan, "Video Script Identification based on Text Lines", In Proc. ICDAR, 2011, pp 1240-1244.
20. M. C. Padma and P. A. Vijaya, "Monothetic Separation of Telugu, Hindi and English Text Lines from Multi Script Document", In Proc. ICSMC, 2009, pp 4870-4875
21. S. Rajkumar and S. Bharathi "Ancient Tamil Script Recognition from Stone Inscriptions Using Slant Removal Method", ICEEBE,2012
22. K. H. Kashyap, Bansilal and P.A. Koushik, "Hybrid Neural Net-work Architecture for age identification of ancient Kannada characters", International Symposium on Circuits and Systems, 2003, Vol.5, pp 661-664.
23. J. Landre, F. M. Nicoier and S. Ruan, "Ornamental letters image classification using local dissimilarity maps", ICDAR, 2009, pp 186-190.
24. V. J. Dongre and V. H. Mankar, "Devanagari document segmentation using histogram approach",IJCSEIT, Vol.1, No.3, August 2011,pp 46-53

25. Rajesh P.N.Rao,"Probablistic Analysis of an Ancient Undeci-phered Script", IEEE Computer Society, Vol. 43(4),April 2010, pp 76-80

26. N. Yadav , M. N. Vahia, I. Mahadevan, H. Joglekar, "Segmentation of Indus texts" , International Journal of Dravidian Linguistics , 2008,Vol. 37(1), pp 53-72.

27. S. K. Sangame, R. J. Ramteke and Y. V. Gundge, "Efficient algorithm for Kannada and English script identification" , ACR,2012,Vol 4, pp 54-56

28. D. Zhao, P. Shivakumara, S. Lu and C. L. Tan, "New Spatial Gradient features for Video script identification", Document Analysis System , 2012,pp 38-42

29. A. S. Kavitha, P. Shivakumara and G. Hemantha Kumar, "An Integrated Method for Classification of Indus and English Document Images", ICERECT, 2012, Karnataka, India (Accepted).

30. T. V. Ashwin and P. S. Sastry, “A font and size-independent OCR system for printed Kannada documents using support vector machines”. *Sadhana,* Vol. 27, No. 1, February 2002, pp. 35-58.

31. B. B. Chaudhuri and U.Pal, “An OCR System to Read Two Indian Languages Scripts: Bangla and Devanagari”. *International Conference on Document Analysis and Recognition,* Aug 1997, pp.1011-1015.

32. B. B. Chaudhuri and U. Pal, “A Complete Printed Bangla OCR System”. *Pattern Recognition*, Vol. 31, No. 5, March 1998, pp.531-539.

33. M. S. Das, D. S. Rani, C. R. K. Reddy and A. Govardhan, “Script identification from Multilingual Telugu, Hindi and English Text Documents”. *International Journal of Wisdom Based Computing* , Vol. 1, No. 3, 2011

34. J.Gllavata and B.Freisleben, “Script Recognition in Images with Complex Backgrounds”. *Signal Processing and Information Technology* , IEEE, 2005, pp. 589-594.

35. M.Grafmuller and J. Beyerer, “Performance improvement of character recognition in industrial applications using prior knowledge for more reliable segmentation”. *Expert Systems with Applications,* Vol. 40, No. 17, 2013, PP. 6955 – 6963.

36. M.Hangargea, K. C. Santoshb, S. Doddamania and R.Pardeshia, “Statistical Texture Features based Handwritten and Printed Text Classification in South

Indian Documents". *International Conference on Emerging Computation & Information Technologies*, Elsevier, 2013, pp. 215-221

37. A. S. Kavitha, P. Shivakumara and G.H. Kumar, "Skewness and Nearest Neighbour Based Approach for Historical Document Classification". *International Conference on Communication Systems and Network Technologies* , IEEE, April 2013, pp. 602-606.
38. P.Krishnan, N. Sankaran, A. K. Singh and C. V.Jawahar, "Towards a robust OCR system for Indic scripts". *Document Analysis Systems* , IEEE, April 2014, pp.141-145.
39. S.Lu, L. Li and and C. L. Tan, "Identification of scripts and orientations of degraded document images". *Pattern Analysis and Applications*, Springer, Vol. 13, No. 4, November 2010, pp.469-475.
40. K. S. Murthy, G. H. Kumar, P.Shivakumara and P. R. Ranganath, "Nearest Neighbour Clustering approach for line and character segmentation in epigraphical scripts". *International Conference on Cognitive Systems,*2004
41. M. C. Padma and P. A. Vijaya "Script identification from trilingual documents using profile based features". *International Journal of Computer Science and Applications*, Vol. 7, No. 4, 2010, pp.16 – 33.
42. J. G. Park and K. J. Kim, "Design of a visual perception model with edge-adaptive Gabor filters and support vector machine for traffic sign detection". *Expert Systems with Applications*, Elsevier, Vol. 40, No.9, July 2013, pp. 3679 – 3687.
43. A.Risnumawan, P. Shivakumara, C. S. Chan and C. L.Tan, "A Robust Arbitrary Text Detection System for Natural Scene Images". *Expert Systems with Applications*, Elsevier, Vol. 41, No 18, pp. 8027-8048.
44. N.Sharma, S.Chanda, U. Pal and M.Blumestiein, "Word-Wise Script Identification from Video Frames". *International Conference on Document Analysis and Recognition*, IEEE, 2013, pp.867-871.
45. P. Shivakumara, Z. Yuan, D. Zhao, T. Lu and C. L.Tan, "New Gradient-Spatial-Structural-Features for Video Script Identification". *Computer Vision and Image Understanding*, Elsevier, Vol. 130, January 2015, PP.35-53.
46. C.Thanuja and G.R. Shreedevi, "Content Based Image Retrieval System for Kannada Query Image from Multilingual Document Image Collection".

International Journal of Engineering Research and Applications, Vol.3, No.4, August 2013, pp. 1329-1335.

47. Iravatham Mahadevan , “Dravidian Proof of the Indus script via the Rig Veda: A Case Study”, BULLETIN OF IRC , No. 4, 2014

48. A.Omar and C. C. Lu, “Text Line Extraction for Historical Document Image using Steerable Directional Filters”, In Proc. ICALIP, 2014, pp312-317.

49. B. Gatos, G. Louloudis and N. Stamatopoulos , “Segmentation of Historical Handwritten Documents into Text Zones and Text Lines”, In Proc. ICFHR, 2014, pp 464-469.

50. F.Kleber, M. Diem and R. Sablatnig, “Robust Skew Estimation of Handwritten and Printed Documents based on Grayvalue Images”, In Proc. ICPR, 2014, pp 3020-3025.

51. A. Garz, A. Fischer , H. Bunke and R. Ingold , “A Binarization-Free Clustering Approach to Segment Curved Text Lines in Historical Manuscripts”, In Proc. ICDAR, 2013,pp 1290-1294.

52. I.Rabaev,O.Biller, J. El-Sana ,K.Kedem and I.Dinstein “Text Line Detection in Corrupted and Damaged Historical Manuscripts”, In Proc. ICDAR, 2013,pp812-816.

53. A. Garz, A. Fischer , R. Sablatnig and H. Bunke , “Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering”, In Proc. DAS, 2012,pp 95-99.

54. I.B. Messaoud, H.Amiri, H.E.Abed and V. Margner, “A Multilevel Text line Segmentation Framework for Handwritten Historical Documents”, In Proc. ICFHR, 2012, pp515-520.

55. A. Soumya and G. H. Kumar, “Preprocessing of Camera Captured Inscriptions and Segmentation of Handwritten Kannada text”, IJARCCE, Vol 3(5), 2014, pp 6794-6803.

56. M. Diem, F. Kleber and R. Sablatnig, “Text Line Detection for Heterogeneous Documents”, In Proc. ICDAR, 2013, pp. 743 - 747.

57. S.S.Bukhari, F.Shafait and T.M. Breuel, “Script-Independent Handwritten Text lines Segmentation Using Active Contours”, In Proc. ICDAR, 2009, pp. 446 - 450.

58. P. Rajesh, P. N. Rao, N. Yadav, M. N. Vahia, Hrishikesh, Joglekar, R. Adhikari, and I. Mahadevan (2010), “Entropy, the Indus Script, and Language”, Computational Linguistics,Vol 36(4),pp 795-805.

59. P. Shivakumara, R. P. Sreedhar, T. Q. Phan, L. Shijian and C. L. Tan, “Multi-Oriented Video Scene Text Detection through Bayesian Classification and Boundary Growing”, IEEE Trans. CSVT, 2012, pp 1227-1235.

60. A. Zhu, G. Wang and Y. Dong, “Robust Text Segmentation in Low Quality Images via Adaptive Stroke Width Estimation and Stroke Based Super pixel Grouping”, Lecture Notes in Computer Science, 2015, pp 119-133.

61. R. Pintus,Y. Yang, H. Rushmeier, “ Automatic Text Height Extraction for the Analysis of text lines in old handwritten manuscripts”, ACM Journal on Computing and Cultural Heritage, Vol. 0, No. 0, 2013, pp 0-25.

62. S. M. Gaurav and C.Nandish, “A - Survey and Analysis of Segmentation, Feature Extraction and Classification in OCR System”, IJAR, Vol 5(1), 2015, pp 24-26.

63. P. Thakur and A.Azam “Edge Detection through Integrated Morphological Gradient and Fuzzy Logic Approach”, IJSETR, Vol 4(5), 2015, pp 1613-1616.

64. A. Cheung, M. Bennamoun, N.W. Bergmann (2001), “An Arabic optical character recognition system using recognition-based segmentation”, Pattern Recognition 34, pp 215-233.

65. A. Choudhary (2014), “A Review of Various Character Segmentation Techniques for Cursive Handwritten Words Recognition”, International Journal of Information & Computation Technology, Vol.4(6), pp. 559-564

66. A. Choudhary, R. Rishi, S. Ahlawat (2013), “A New Character Segmentation Approach for Off-Line Cursive Handwritten Words”, Information Technology and Quantitative Management, Vol 17, pp 88 – 95.

67. A. Madhavaraj, A.G. Ramakrishnan, H.R. Shiva Kumar, N. Bhat (2014), " Improved recognition of aged Kannada documents by effective segmentation of merged characters", In Proc. SPCOM, pp 1-6.

68. A. R. Khan & D. Muhammad (2008), "A Simple Segmentation Approach for Unconstrained Cursive Handwritten Words in Conjunction with the Neural Network", International Journal of Image Processing, Vol 2(3), pp 29-35.

69. A.Choudhary, S. Ahlawat, R.Rishi (2015), "A Neural Approach to Cursive Handwritten Character Recognition Using Features Extracted from Binarization Technique", Computations Studies Volume 319, pp 745-771.

70. C. Naveena, V.N. M. Aradhya (2012), "Handwritten Character Segmentation for Kannada Scripts", In Proc. IICT, pp. 144-149.

71. C. Silva and C. Kariyawasam (2014), "Segmenting Sinhala Handwritten Characters", International Journal of Conceptions on Computing and Information Technology Vol. 2(4).

72. G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis (2008), "A Complete Optical Character Recognition Methodology for Historical Documents", In Proc. DAS, pp 525-532.

73. I.Supriana, A. Nasution (2013), "Arabic Character Recognition System Development", In Proc. ICEEI, pp. 334 – 341.

74. N. Anupama, Ch. Rupa& E.S. Reddy (2013), "Character Segmentation for Telugu Image Document using Multiple Histogram Projections", Global Journal of Computer Science and Technology Graphics &Vision,Vol 13(5).

75. N. Sharama, P. Shivakumara, U.Pal, M.Blumenstien and C. L. Tan (2013), "A New Method for Character Segmentation from Multi-Oriented Video words", In Proc. ICDAR, pp 413-417.

76. N.Nikolaou, M.Makridis , B. Gatos , Nikolaos Stamatopoulos , Nikos Papamarkos(2010), "Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation path", Image and Vision Computing ,Vol 28(4).

77. N.Sridevi and P.Subashini (2012), "Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques", IJCA, Vol 52(4), pp 7-12.

78. P. Mathivanan, B. Ganesamoorthy2 and P. Maran3 (2014), "Watershed algorithm based segmentation for handwritten text identification", ICTACT Journal on Image and Video processing, Vol 4(3).

79. P. Rajesh and P. N. Rao (2010), "Probabilistic Analysis of an Ancient Undeciphered Script", In Proc. Computer Society, pp 76-80.

80. S. Kapoor and V.Verma (2014), "Fragmentation of Handwritten Touching characters in Devanagari Script", International Journal of Information Technology, Modeling and Computing, Vol. 2(1).

81. S.A. Angadi and M.M. Kodabagi(2014), "A Robust Segmentation Technique for Line, Word and Character Extraction from Kannada Text in Low Resolution Display Board Images" , Signal and Image Processing , 2014, pp 42-49.

82. T. Saba, A. Rehman, S. Al-Zahrani, (2014) "Character Segmentation in Overlapped Script using Benchmark Database", Computers, Automatic Control, Signal Processing and Systems Science.

83. T. V. Phan , B. Zhu , M. Nakagawa (2011), "Development of Nom Character Segmentation for Collecting Patterns from Historical Document Pages" ,In Proc. HDIP, pp 133-139.

84. Y. Mei, X. Wang, J. Wang (2013),"An Efficient Character Segmentation Algorithm for Printed Chinese Documents", ACIT 2013, ASTL Vol. 22 , pp. 183 - 189

85. Chacko A. M. M. O and Dhanya P. M (2015), "A Comparative Study of Different Feature Extraction Techniques for Offline Malayalam Character Recognition", Computational Intelligence in Data Mining, pp. 9-18.

86. Das S and Banerjee S (2014), "Survey Of Pattern Recognition Approaches in Japanese Character Recognition", International Journal of Computer Science and Information Technologies, pp. 93- 99.

87. Das S and Banerjee S (2015), "An Algorithm for Japanese Character Recognition", Image, Graphics and Signal Processing, pp. 9- 15.

88. Dave N (2015), "Segmentation Methods for Hand Written Character Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition, 2015, pp. 155-164.

89. De-Campos T, Babu B and Verma M (2009), "Character recognition in natural images", In Proc. VISAPP.

90. Dholakia K (2015), "A Survey on Handwritten Character Recognition Techniques for Various Indian Languages ", International Journal of Computer Applications, pp. 17-21.

91. Farhad M. M, Hossain S. M. N, Khan A. S, and Islam A (2014),"An Efficient Optical Character Recognition Algorithm using Artificial Neural Network by Curvature Properties of Characters", International Conference On Informatics, Electronics & Vision, pp. 1-5.

92. Hegadi R. S Kamble P. M (2014), "Recognition of marathi handwritten numerals using multi-layer feed-forward neural network", In World Congress on Computing and Communication Technologies, pp. 21–24.

93. Hirabara L. Y, Aires S. B. K, Freitas S. C. O. A, Britto A. S and Sabourin R (2011), "Dynamic Zoning Selection for Handwritten Character Recognition ", In Proc. Image Analysis and Compute Vision Applications, pp. 507–514.

94. Honarvar B, Paramesran R and Lim C. L (2014), "Image reconstruction from a complete set of geometric and complex moments", Signal Processing, pp. 224-232.

95. Hussain E, Hannan A and Kashyap K (2015), "A Zoning based Feature Extraction method for Recognition of Handwritten Assamese Characters", International Journal of Computer Science And Technology, pp 226-228.

96. Iamsa-at S and Horata P (2013), "Handwritten Character Recognition Using Histograms of Oriented Gradient Features in Deep Learning of Artificial Neural Network," International Conference in IT Convergence and Security, pp. 1- 5.

97. Ismail S. M and Abdullah S. H. S (2013), "Geometrical-Matrix Feature Extraction for On-Line Handwritten Characters Recognition", Journal of Theoretical and Applied Information Technology, pp. 86- 93.

98. Kale K. V, Deshmukh P. D, Chavan S. V, Kazi M. M and Rode Y. S (2014), "Zernike Moment Feature Extraction for Handwritten Devanagari (Marathi) Compound Character Recognition", International Journal of Advanced Research in Artificial Intelligence, pp. 68-76.

99. Kamblea P. M and Hegadi R. S (2015), "Handwritten Marathi character recognition using R-HOG Feature", International Conference on Advanced Computing Technologies and Applications, pp. 266–274.

100.Karatzas D, Mestre S. R, Mas J, Nourbakhsh F and Roy P. P (2011), “ICDAR 2011 Robust Reading Competition”, In Proc. ICDAR, pp 1485-1490.

101. Kaushal D. S, Khan Y and Varma S (2014), “Handwritten Urdu Character Recognition Using Zernike Feature Extraction and Support Vector Machine Classifier”, International Journal of Research, pp.1084-1089.

102. Kavitha A. S, Shivakumara P and Kumar G. H (2016), “Text Segmentation in degraded historical document Images”, Egyptian Informatics Journal, 2016.

103. Kavitha A. S, Shivakumara P, and Kumar G. H (2015), “A Robust Script Identification System for Historical Indian Document Images”, Malyasian Journal of Computer Science, pp 283-300.

104. Khare V, Shivakumara P and Raveendran P (2015), “A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video”, Expert Systems with Applications, pp 7627-7640.

105. Lawgali, A (2015), “A Survey on Arabic Character Recognition”, International Journal of Signal Processing, Image Processing and Pattern Recognition, pp. 401-426.

106. Li B, Peng L and Ji J (2014), “Historical Chinese Character Recognition Method Based on Style Transfer Mapping”, International Workshop on Document Analysis Systems, pp. 96-100.

107. Matsuda T, Iwamura M and Kise K (2014), “Performance Improvement in Local Feature Based Camera-Captured Character Recognition”, International Workshop on Document Analysis Systems, pp. 196-201.

108. Nemmour H and Chibani Y (2011), “Training Tangent Similarities with N-SVM for Alphanumeric Character Recognition,” American Journal of Signal Processing, pp.34-39.

109. Nguyen P. X, Wang K and Belongie S (2014), “Video Text Detection and Recognition: Dataset and Benchmark”, In Proc. WACV, pp 776-783.

110. Pawar V. R and Gaikwad A (2014), "Multistage Recognition Approach for Offline Handwritten Marathi Script Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition, pp.365-378.

111. Rodrguez K. C. O, Chavez G. C and Menotti D (2012), “Hu and Zernike Moments for Sign Language Recognition”, International Conference on Image Processing, Computer Vision, and Pattern Recognition, pp1-5.

112. Roy S, Shivakumara P, Roy P. P, Pal U and Tan C. L (2015), “Bayesian classifier for multi-oriented video text recognition system”, Expert Systems with Applications, pp 5554-5566.

113. Sadanand A. K, Prashant L. B, Ramesh R. M and Yannawar L. P (2015), “Offline MODI Character Recognition Using Complex Moments”, International Symposium on Computer Vision and the Internet, pp. 516 – 523.

114. Sahlol A. T, Suen C. Y, Elbasyouni M. R and Sallam A. A (2014),"A Proposed OCR Algorithm for the Recognition of Handwritten Arabic Characters", Journal of Pattern Recognition and Intelligent Systems, pp. 8-22.

115. Saleem S, Hollaus F, Diem M and Sablatnig R (2014), “Recognizing Glagolitic Characters in Degraded Historical Documents”, In Proc. International Conference on Frontiers Handwriting Recognition, pp 771-776.

116. Salouan R, Safi S and Bouikhalene B (2014), “A Comparative Study between the Pseudo Zernike and Krawtchouk Invariants Moments for Printed Arabic Characters Recognition”, Journal of Emerging Tech. in Web Intelligence, pp. 89-93.

117. Schwenk H and Milgram M (1995), “Learning Discriminant Tangent Models forHandwritten Character Recognition,” International Conference on Artificial Neural Networks (ICANN), pp. 585–590.

118. Sedighi A and Vafadust M (2011), “A new and robust method for character segmentation and recognition in license plates images”, Expert Systems with Applications, pp 13497-13504.

119. Sikdar A, Banerjee S, Roy P, Mukherjee S and Das M (2014), “Bengali printed character recognition using a feature based chain code method”, Advances in Image and Video Processing, pp. 01–09.

120. Simard P. Y, Cun Y. A. L, Denker J. S and Victorri B (2000), “Transformation Invariance in Pattern Recognition: Tangent Distance and Propagation,” John Wiley & Sons, Inc, pp.181-197.

121. Surinta O, Karaaba M. F, Schomaker L. R. B and Wiering M. A (2015), “Recognition of handwritten characters using local gradient feature descriptors”, Engineering Applications of Artificial Intelligence, pp. 405–414.

122. Tikader A and Puhan N. B (2014), “Edge based directional features for English-Bengali Script Recognition”, International Conference on Signal Processing and Integrated Networks.

123. Wahi A, Sundaramurthy S, and Ponnusamy P (2014), “A comparative study for handwritten Tamil character recognition using wavelet transform and Zernike moments,” International Journal of Open Information Technologies, pp. 30–35.

124. Yang W, Jin L, Xie Z and Feng Z (2015), “Improved deep convolutional neural network for online handwritten Chinese character recognition using domain-specific knowledge”, Document Analysis and Recognition,pp. 551 – 555.

http://en.wikipedia.org/wiki/Indus scripts

http://en.wikipedia.org/wiki/Brahmic scripts